Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), W1-W33. https://doi.org/10.7326/M18-1377

PROBAST: a tool to assess risk of bias and applicability of prediction model studies – explanation and elaboration

Prof Karel G. M. Moons, PhD[1,2,*], Robert F. Wolff, MD[3,*], Prof Richard D. Riley, PhD[4], Penny F. Whiting, PhD[5,6], Marie Westwood, PhD[3], Prof Gary S. Collins, PhD[7], Prof Johannes B. Reitsma, MD, PhD[1,2], Prof Jos Kleijnen, MD, PhD[3,8], Sue Mallett, DPhil[9]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[2] Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands

[3] Kleijnen Systematic Reviews Ltd, York, United Kingdom

[4] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, Staffordshire, United Kingdom

[5] Bristol Medical School, University of Bristol, Bristol, United Kingdom

[6] NIIHR CLAHRC West, University Hospitals Bristol NHS Foundation Trust, Bristol, United Kingdom

[7] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Diseases, University of Oxford, Oxford, United Kingdom

[8] School for Public Health and Primary Care (CAPHRI) Maastricht University, Maastricht, The Netherlands

[9] Institute of Applied Health Research, NIHR Birmingham Biomedical Research Centre, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[*] Both authors contributed equally

Corresponding author:
Prof Karel G. M. Moons
Julius Centre for Health Sciences and Primary Care
UMC Utrecht
PO Box 85500
3508 GA Utrecht
The Netherlands
K.G.M.Moons@umcutrecht.nl

Short title: PROBAST E&E
Word count: 17,631 words
(Introduction, Focus of PROBAST, Risk of bias and applicability, Applying PROBAST, Concluding remarks)
Keywords: Bias (Epidemiology); Diagnosis, Evidence-Based Medicine; Multivariable Analysis; Prediction; Prognosis; Reproducibility of Results

Abstract

(230 words)

Prediction models in healthcare aim to estimate for an individual, the probability that a condition or disease is already present (diagnostic model) or that an outcome will occur in the future (prognostic model), based on multiple predictors.

Publications on prediction models have increased in recent years, and there are often competing prediction models for the same outcome or target population. Healthcare providers, guideline developers and policymakers are often unsure which model to use or recommend, and in which individuals or settings. Hence systematic reviews of these studies are increasingly demanded, required and performed.

A key part of a systematic review of prediction models is to examine the risk of bias and applicability for the intended population. To help reviewers with this process, we developed PROBAST, a Prediction model Risk Of Bias ASsessment Tool for studies developing, validating or updated (e.g. extending) prediction models, both diagnostic and prognostic, models.

PROBAST was developed through a consensus process involving a group of experts in the field. PROBAST includes four domains (Participants; Predictors; Outcome; Analysis) containing 20 signalling questions. This Explanation and Elaboration paper describes the rationale for including each domain and signalling question and provides guidance for reviewers on how to use these to assess risk of bias and applicability concerns. All concepts are illustrated with published examples across different topics. The PROBAST checklist and accompanying documents can also be downloaded from www.probast.org.

Introduction

(532 words)

Prediction models in healthcare often aim to predict for an individual, whether a particular outcome such as disease is present (diagnostic models) or will occur in the future (prognostic models).(1-6) Diagnostic models can be used to refer patients for further testing, to initiate treatment or to inform patients. Prognostic models can be used for decisions on preventive lifestyle changes, therapeutic interventions or monitoring strategies, or for risk stratification in randomised trial design and analysis.(7, 8) Potential users of prediction models include healthcare professionals, policy makers, guideline developers, patients and the general public.

In the medical literature, there are thousands of studies developing and validating prediction models and often numerous prediction models for the same target population and outcomes. For example, there are over 60 models for breast cancer prognosis,(9) over 250 models in obstetrics,(10) and nearly 800 models predicting outcomes in patients with cardiovascular disease.(11) This growth of prediction models will increase further with the growth of personalized or precision medicine.

Systematic reviews are the most reliable form of evidence for decision makers for randomised therapeutic studies and diagnostic test accuracy studies.(12) There is growing interest in systematic reviews of prediction model studies, as exemplified by the formation of the Cochrane Prognosis Methods Group to support systematic reviews of prognosis, including prognostic model studies.(13, 14) Guidance to facilitate systematic reviews of prediction models has been developed (Table 1) including for search strategies(15-18), formulating the review question (14+15), data extraction(19)), and meta-analysis (20-22).

Assessment of the risk of bias (ROB) is an essential step in any systematic review. Shortcomings in study design, conduct and analysis can result in study estimates being at 'risk of bias', i.e. flawed or distorted results. When interpreting results from a systematic review, stronger conclusions can be drawn from a systematic review based on primary studies at low ROB rather than studies at high or unclear ROB.(44) It is also important to identify the studies with most relevance to the settings and populations targeted in the review, based on the applicability of primary studies for the review question. We developed PROBAST (Prediction model Risk Of Bias ASsessment Tool) due to the lack of suitable tools designed specifically to assess risk of bias and applicability of primary prediction model studies.

PROBAST consists of four domains, with 20 signalling questions to facilitate ROB assessment.(REF M18-1376] The structure and rating is similar to tools designed to assess the risk of bias in randomised trials (revised Cochrane tool, ROB 2.0), diagnostic accuracy studies (QUADAS-2) and systematic reviews (ROBIS).(42, 45, 46) Although PROBAST is initially designed for use in systematic reviews of prediction model studies, it can also be used as a general critical appraisal tool for prediction model studies.

Here we describe the rationale behind the domains and signalling questions, how to use them, and how to reach domain level and overall judgements on risk of bias and applicability of primary studies to the review question. We illustrate using examples from across the medical field using six filled-in examples. As this is an area of active research, the PROBAST tool, examples and accompanying guidance will be updated when needed, and the latest PROBAST tool version should always be downloaded from the website (www.probast.org).

Focus of PROBAST

(954 words)

PROBAST is designed to assess primary studies that developed, validated, or updated (e.g. extended) one or more multivariable prediction models for diagnosis or prognosis (Boxes 1 and 2). A multivariable prediction model is defined as any combination or equation of two or more predictors (e.g. age, gender, disease stage, biomarkers) for estimating the probability or risk for an individual.(1, 4, 6-8, 47-49) Other names for prediction model include risk prediction model, predictive model, prediction index or rule, and risk score.(1, 3-8, 49, 50)

Diagnostic and prognostic models

Diagnostic prediction models estimate the probability that a certain outcome, the "target condition", is currently present. Diagnostic prediction model studies typically include individuals who are *suspected of having* the target condition but not yet known to *have* it.

Prognostic prediction models estimate the probability that an outcome or event will occur, e.g. death, disease recurrence, disease complication, or therapy response. The time period of prediction can vary from hours, e.g. pre-operatively predicting post-operative nausea and vomiting, to years, e.g. predicting life-long risk of developing a coronary event. Although many prognostic models enrol patients with an established diagnosis, this does not have to be the starting point, as seen in models for predicting the development of diabetes in pregnant women(51) or of osteoporotic fractures in the general population(52). Consistent with the TRIPOD statement(7, 8), PROBAST thus uses a broad definition of prognostic models referring to the prediction of future outcomes, studied in individuals at risk of developing that outcome.

Diagnostic and prognostic model studies often use different terms for predictors and outcomes (Box 2). In the cancer literature, often a distinction is made between prognostic versus predictive models, where predictive models refer to identifying individuals with differential treatment effects.(53) For this manuscript, these types of (predictive) models are out of scope.

Types of predictors, outcomes and modelling technique

PROBAST can be used to assess any type of diagnostic or prognostic prediction model aimed at individualised predictions, regardless of the predictors used, outcomes being predicted, or method to develop, validate or adjust the model.

Predictors range from demographics, medical history and physical examination to results from imaging, electrophysiology, blood and urine measurements, pathological examinations, disease stages or characteristics, to results from -omics and any new biological measurement. Predictors are also referred to as covariates, risk indicators, prognostic factors, determinants, index test results or independent variables.(4, 6-8, 49, 54, 55)

PROBAST distinguishes between candidate predictors and predictors included in the final model.(56) Candidate predictors are those variables considered to be potentially predictive of the outcome, i.e. all those evaluated in the study whether or not included in the final multivariable model.

PROBAST primarily addresses prediction models for binary and time-to-event outcomes, as these are the most common in medicine. However, PROBAST can also be used to assess models predicting non-binary outcomes such as continuous scores, for example pain scores or cholesterol levels, or

categorical outcomes such as the Glasgow Coma scale. Almost all PROBAST signalling questions apply equally to the assessment of prediction models for continuous and categorical outcomes, except signalling questions addressing number of outcome events per predictor, and certain model performance measures (e.g. c-statistic), which are not relevant to continuous outcomes.

Prediction models usually involve regression modelling techniques such as logistic regression or survival models. Prediction models may also be developed or validated using non-regression techniques such as neural networks, random forests or support vector machines. As the use of routine big data increases, additional modelling techniques are becoming more common, such as machine and artificial learning models. The main differences between studies using regression and other types of prediction modelling include the methods of data analysis; non-regression development models can often have greater risks of overfitting when data are sparse, and the potential lack of transparency can affect the applicability and usability of the models.(57) Below we provide guidance how PROBAST can be adapted to address other types of outcomes and modelling techniques.

Types of review question

PROBAST can be used to assess different types of systematic review questions. For some review questions it is relevant to include all prediction model studies including both development and validation, but for other questions only validation studies would be relevant. Box 3 gives examples of potential review questions for both prognostic and diagnostic prediction models where PROBAST is applicable. The CHARMS Checklist provides explicit guidance on how to frame a focused question for reviews of prediction model studies.(19)(20)

Types of prediction model studies

PROBAST addresses studies on multivariable models that are to be used to make predictions in individuals, i.e. *individualised predictions* (Box 1), including studies on:

- development of new prediction models
- development and validation of the same prediction model(s)
- validation existing prediction models
- development of new compared with validation of existing prediction models
- updating (e.g. adjusting model coefficients) or extension (e.g. adding new predictors) of existing prediction models
- combinations of the above.

PROBAST is not designed for assessing predictor finding studies where the aim of multivariable modelling is to identify predictors associated with outcome, rather than developing a model for individualised predictions.(19, 68, 69); the QUIPS tool has been developed for assessment of bias in these studies.(70)

PROBAST is also not suitable for assessing comparative studies that quantify the impact on participants' health outcomes of using a prediction model (as part of a complex intervention) in comparison to not using a model or an alternative (Box 1). Such comparative model impact studies use either randomised or non-randomised designs(71-74) and appropriate risk of bias tools for randomised studies (45) or non-randomised studies(75).

For diagnostic test accuracy studies, another ROB tool, QUADAS-2, has been developed.(46) However, it should be noted that some diagnostic test accuracy studies include a diagnostic prediction model. In these cases, the use of PROBAST should be considered.

Risk of bias and Applicability

(335 words)

Risk of Bias

Bias is usually defined as presence of systematic error within a study leading to distorted or flawed study results, hampering the internal validity of that study. In prediction model development and validation, there are known features which make a study at risk of bias, although there is limited *empirical* evidence to demonstrate the most important sources of bias. We define risk of bias to occur when shortcomings in study design, conduct or analysis lead to systematically distorted estimates of model predictive performance. Model predictive performance is typically evaluated using measures pf calibration and discrimination, and sometimes (notably in diagnostic model studies) classification (Box 4).(8) When assessing risk of bias, it helps to think about how the equivalent hypothetical methodologically robust prediction model study would have been designed, conducted and analysed.

Applicability

Concerns for the applicability of primary studies to the review question can arise when the study population, predictors or outcomes of a primary study differ from those specified in the review question. For example, applicability concerns may arise when participants in the prediction model study are from a different medical setting than the targeted population defined in the review question (Table 2). A prediction model developed in secondary care may have different discrimination and calibration in primary care as patients in hospital settings typically have more severe disease than patients in primary care.(71, 86)

For systematic reviews where participants, predictors and outcomes of the primary studies directly match the review question, there will likely be small concerns about applicability of the study. However, typically systematic reviews have inclusion criteria that are broader than the precise focus of the review question.

We note that bias and applicability concerns should here not be confused with heterogeneity in predictive performance of a particular model across different validation studies, that may result for example from different disease severities or case-mix.(21) Variation of performance of a model across multiple validations can be reported with relevant prediction intervals, as part of investigation of heterogeneity using meta-analysis methods.(20)

Applying PROBAST

(15,502 words)

The PROBAST tool consists of four steps (Table 3). A PROBAST assessment should be completed for each distinct model that is relevant to the systematic review question. We use a variety of examples to illustrate key issues relating to risk of bias and applicability (Table 4). These examples address diagnostic and prognostic models, different medical areas, study designs, predictor and outcome types, and include development and validation studies. Assessments of these examples are available at www.probast.org.

Step 1 – Specify your review question(s)

First reviewers need to specify their review question in terms of intended use of the prediction model, targeted participants, predictors used in the modelling, and outcomes to be predicted. Structured reporting of these elements facilitates assessment of applicability. Specific guidance (i.e. the CHARMS checklist) exists to help reviewers define a clear and focused review question (19), summarized in Table 2.

Step 1 is completed once per systematic review. Table 5 provides an example.

Step 2 – Classify the type of prediction model evaluation

In Step 2 the type of prediction model evaluation is identified to link to the relevant signalling questions in PROBAST. When both, development and validation (see Box 1) of a particular model, is of interest and reported in a single publication, each will be assessed separately. Similarly, when a certain model is being validated and adjusted or extended in the same publication. A model extension, where new predictors are added to an existing model, would be assessed as new model development.

Step 2 is completed once for each prediction model assessed for the review (Table 6 provides an example).

Step 3 – Assess risk of bias and applicability

Assessing risk of bias

PROBAST provides a structured approach to identify potential risk of bias, based on four domains with signalling questions. Signalling questions are *factual* questions and are rated as yes (Y), probably yes (PY), no (N), probably no (PN), or no information (NI). All signalling questions are phrased so that "yes" indicates low risk of bias, and "no" high risk of bias. The ratings of PY and PN are included to allow judgements to be made when there is *not sufficient* information for reviewers to be confident of making a Y or N rating. Conforming to other risk of bias tools, responses of "yes" are intended to have similar implications to responses of "probably yes" (and similarly for "no" and "probably no"), but allow for a distinction between something that is known and something that is likely to be the case.(42, 45, 75) "No information" should only be used when there is truly no information to answer a signalling question.

The answers to these signalling questions assist reviewers when judging the overall risk of bias for each domain. A domain where all signalling questions are answered Y or PY should be judged as "low risk of bias". An answer of N or PN on one or more signalling question flags the potential for bias while NI indicates insufficient information. This does not mean that bias is definitely present. For example, in a prognostic study where predictors were clearly determined before event occurrence and

measurement, but the report does not state whether predictor measurements were blinded for information on the outcome occurrence, this signalling question (2.3, see below) is factually rated as NI. However, in the overall risk of bias judgement of this domain one may still judge it to be low risk of bias, since it can be inferred that predictors were measured a long time before the outcome occurred. When judging risk of bias for a particular domain, reviewers thus need to use their judgement to determine whether or not issues identified by the signalling questions are likely to have introduced bias into the model development or validation.

Assessing concerns for applicability

Applicability of a primary study to match the review question is assessed for the first three domains using information reported in Table 5 (the review question) and Tables 7 to 9. The analysis domain relates to limitations with the data or how the analysis was performed, which are not related to the review question, and so has no applicability assessment. The degree of applicability is rated as "low", "high" or "unclear" concern. The "unclear" category should only be used when insufficient information is reported.

If there is a good match between the review question and the primary study, there are likely to be low concerns concerning applicability. Often, a review may address a focused question but study inclusion criteria are set broader.

Support for judgement and rationale for rating

To improve the transparency of the assessment process, PROBAST includes two types of text boxes for each domain. The first "support for judgement" box, allows reviewers to record information that was used to answer the risk of bias signalling questions or inform the applicability assessment for that domain. Text can either be copied and pasted directly from the article being assessed, or summarised. The second text box is the "rationale for rating" allowing reviewers to record the reason for judging the model at high, low or unclear risk of bias or having high, low or unclear concerns for applicability, respectively. For example, if a domain is judged at high risk of bias, the reviewers can summarise which study features led to the rating. Or, if a domain is rated as low risk of bias despite one or more signalling questions being rated as "no", "probably no" or "no information", this box can be used to explain why issues identified by the signalling questions are not likely to have introduced bias into the study.

Further guidance and examples are provided in the relevant domain specific sections as well as Tables 7 to 10. Latest updated versions of guidance can be downloaded from www.probast.org.

Domain 1: Participants

This domain covers potential sources of bias and applicability concerns related to how participants were selected for enrolment into the study and the data sources used. In the support for judgement box, reviewers should describe the sources of data that were used, for example from a cohort study, randomised study, or routine care registry, and the criteria for participant selection in the primary study.

Risk of bias

There are two signalling questions to facilitate risk of bias judgment for this domain (Table 7).

1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?

Numerous data sources or study designs can be used in prediction model studies.

Prognostic model studies

Prognostic model studies are at a low risk of bias when based on a prospective longitudinal cohort design, where methods tend to be defined and consistently applied for participant inclusion and exclusion criteria, predictor assessment and for outcome determination across a predefined follow up.(1) Using pre-specified and consistent methods ensures that the participants and related data are systematically and validly recorded.

The potential for risk of bias in model development and validation studies is increased when participant data are from existing data sources, such as data from existing cohort studies or routine care registries. This is because data are often not collected using a protocol that was designed specifically for prediction model purposes but for some other purpose. For routine care registries, data relating to inclusion and exclusion criteria are often inconsistently measured and recorded.(21, 91) For example, in relation to the Clinical Practice Research Datalink (CPRD), Herrett et al. state that "the quality of primary care data is variable because data are entered by GPs during routine consultations, not for the purpose of research. Researchers must therefore undertake comprehensive data quality checks before undertaking a study".(91)

Data from one or more arms of randomised intervention trials can also be used for prognostic model development or validation. However, the randomised treatments may need to be included as separate predictors to account for any treatment effects, as effective treatments are predictors of the outcome.(92, 93) RCTs also usually have more restricted inclusion criteria typically leading to smaller distributions of the predictors (so-called smaller case-mix). It has been shown that models developed or validated using data with smaller predictor distribution (smaller case mix) tend to show a lower discriminative ability than models developed or validated from data sources where the predictors have a broader distribution.(94-97) This is because in the former the range of a model's predicted probability is smaller and therefore the discriminative ability of the model is smaller as well.

Case-cohort or nested case-control studies, in which participants with the outcome (cases) and without the outcome (non-cases or controls) are sampled from a pre-existing, well described cohorts or routine care registries of known size, can be considered at low risk of bias provided researchers appropriately adjust for the original cohort or registry outcome frequency in the analysis (see signalling question 4.6).(56, 98-101) If they do not, the study is at high risk of bias. For example, for logistic prediction models, reweighting the controls and cases by the inverse sampling fraction (from the original cohort or registry) allows correct estimation of baseline risk, allowing corrected absolute predicted probabilities and model calibration measures to be obtained.(98-101) Case-control studies in which cases and controls are not sampled from a pre-specified and well defined cohort or registry, are at high risk of bias. This is because the definition and number of the selected cases and controls relative to the source population is unclear. Accordingly, baseline risks or hazards and absolute outcome probabilities cannot be correctly adjusted for.(56)

Diagnostic model studies

Diagnostic models predict the presence or absence of an outcome (target disease) at the same time point as the index tests or predictors are measured (Box 2). Accordingly, the design with lowest risk of bias for diagnostic model studies is a cross-sectional study where a group (cohort) of participants is selected based on having certain symptoms or signs that makes them 'suspected of having the target condition of interest'. Subsequently, the predictors (index tests) and outcome (disease presence or absence) according to the reference standard are measured in all participants.(102-105) Diagnostic

studies using a cross-sectional design in which the presence of disease cannot be determined in all patients by the reference standard in all patients (e.g. some participants with potential malignant mass have no lesion on imaging which can be biopsied), require additional follow-up of participants over time to establish whether the target condition was present when the index tests were performed.

As with prognostic models, a diagnostic model using a nested-case-control design can only be at low risk of bias if researchers adjust the case and control samples by the inverse sampling fractions (see signalling question 4.6) to obtain correct estimate of the outcome prevalence in the original cohort.(106-110) Similarly, if a non-nested case control design is used, where advanced cases and healthy controls are over-represented, this will lead to incorrect estimates of disease prevalence and overestimated diagnostic model performance.(107-110)

> **Example:**
> In Perel 2012, data for the development of the prognostic model came from a randomised trial (CRASH-2), combining the data from the two treatment arms.(89) As the authors included the allocated treatment as a predictor in the prediction model development, this signalling question should be answered as Y.
> Aslibekyan 2011 used a non-nested case-control study but the authors did not adjust their analyses by weighting the cases and controls by the inverse of the sampling fractions.(87) Accordingly, this signalling question for this study should be answered as N.

1.2 Were all inclusions and exclusions of participants appropriate?

Studies that make inappropriate inclusions or exclusions of study participants may result in biased estimates of model predictive performance as the model is based on a selected subgroup of participants that may not be representative of the intended target population.

Inappropriate inclusion results from including participants already known to have the outcome at the time of predictor measurement. For example, in a study developing a model to predict the future development of type II diabetes, some participants may already have type II diabetes if study inclusion criteria were based on participants without diabetes solely using self-reported criteria. Including participants who already have diabetes will most likely result in a model with overestimated predictive performance.

Similarly, for a diagnostic model that aims to detect the presence or absence of pulmonary embolism in symptomatic patients, the exclusion of patients with pre-existing lung disease could be considered an example of an inappropriate exclusion. Patients with pre-existing lung disease may be harder to diagnose with pulmonary embolism than those without pre-existing lung disease; diagnostic accuracy may be overestimated if a model, after excluding these patients, is developed for use in all patients suspected of pulmonary embolism. Authors should then explicitly state that the developed model is only applicable to suspected lung embolism patients without pre-existing lung disease.

Note that this signalling question is not asking about loss to follow up of participants after inclusion in the primary study (i.e. it is not about inappropriate exclusions during the study); this is dealt with in domain 4. This signalling question is about participants who were inappropriately included or excluded from the study. Further, it is important to distinguish between a selection bias imposed on a study population by restrictions in inclusion criteria, compared to a study population with different characteristics that may limit the applicability of the study to the review question (see below under applicability).

In summary, the key issue is whether any inclusion or exclusion criteria, or the recruitment strategy, could have made the included study participants unrepresentative of the intended target population for the review. Some risk of bias tools (e.g. QUADAS-2) have a signalling question asking whether the study recruited a consecutive or random sample of patients. As this is rarely achievable for any study, we have not included this as a signalling question in PROBAST.

> **Example:**
> Aslibekyan et al. excluded all participants with a fatal myocardial infarction (MI) because they used a case-control design.(87) Participants who had died of fatal-MI were excluded as retrospective self-reported data could not be collected from these patients. The prediction model for non-fatal MI was thus based on selected healthier participants, including only those who survived an MI or did not develop a MI (controls). This is likely to have introduced bias as the study participants represent a selected 'lower-risk-sample' of the original 'at risk of MI population'. Stating that the developed prediction model only predicts non-fatal MI does not solve the issue since at the moment of prediction it is not possible to identify participants who will develop fatal-MI, i.e. this signalling question should be answered as PN.

Rating the risk of bias for domain 1

Table 7 shows how the signalling questions should be answered and an overall judgement for domain 1 reached.

Applicability

Applicability for this domain considers the extent to which the population included in the primary study matches the participants specified in the systematic review question (step 1, Table 5). Consider a review with the aim of identifying all model development and validation studies to diagnose bacterial conjunctivitis in symptomatic children. The review could specify inclusion criteria such that prediction model studies with both, adults and children, were eligible. Studies that included only children would be likely to receive a rating of low concern for applicability, whereas studies conducted in adults and children may be rated as at high concern for applicability.

The generalisability and thus applicability of prediction model studies based on randomised trial data needs careful consideration. Randomised trials tend to apply strict inclusion and exclusion criteria, may measure fewer predictors and outcomes, thus reducing the applicability of a model developed or validated from trial data. In contrast, distribution of study characteristics, predictors and outcomes, and thus the generalisability of prediction model studies tends to be high when data from routine care or health care registries are used for model development or validation.

It is often challenging to identify when certain issues relating to a primary study are likely to introduce risk of bias or whether these are concerns for applicability. Applicability assessment is entirely dependent on the systematic review question. Consider the hypothetical pulmonary embolism example in signalling question 1.2 where reviewers might restrict the intended target population of their review, to 'patients suspected of having pulmonary embolism without pre-existing lung disease'. For this target population, a primary study including patients with pre-existing lung disease would constitute an applicability concern and not necessarily a risk of bias. Similarly, consider a diagnostic model development study that included patients with a broad age range (18 to 90 years). This may not have introduced any bias into the primary study but it may limit the applicability of the model if the systematic review question focuses on young adults only (18 to 30 years).

Finally, in a review and meta-analysis of a specific single model, that includes all validation studies of that model, risk of bias and applicability assessments should be supplemented with an investigation of heterogeneity in the reported predictive performance of that model across the validation studies. The

predictive performance of a specific model validated in other studies, is expected to be different due to differences in for example participant characteristics, healthcare setting, geographical location or calendar time periods. This does not mean there is risk of bias *within* the primary validation study or there are concerns about applicability; it merely reflects expected variation in predictive performance of a specific model across studies. Potential sources of heterogeneity between studies can be investigated using meta-analysis or presentation stratified by characteristics that differ across studies.(20, 21)

Also note that sometimes studies validate a model that was developed in a specific group of participants, i.e. in participant data that were (for the researchers) *intentionally* different from the development study. For example, models developed from a healthy general population to predict cardiovascular outcomes, have been validated in patients diagnosed with type II diabetes mellitus.(111) Another example is validating the diagnostic performance of a model to diagnose deep vein thrombosis that was developed in an emergency secondary care setting in a primary care setting.(86) In both cases, heterogeneity in model performance between the development study and the validation studies should be expected.

Domain 2: Predictors

This domain covers potential sources of bias and applicability concerns related to the definition and measurement of the predictors. Predictors are the variables evaluated for their association with the outcome of interest, and ultimately included in combination to form the the prediction model.

In the support for judgement box reviewers may list and describe how the predictors were defined, the time point of their assessment and whether other information was available when assessing the predictors.

Note that for systematic reviews focusing on a specific prediction model, it is sufficient to list and describe only the predictors in the model being validated.

Risk of bias

There are three signalling questions to facilitate a risk of bias judgment for this domain (Table 8).

2.1 Were predictors defined and assessed in a similar way for all participants?

Predictors should be defined and assessed in the same way for all study participants to reduce risk of bias. If different definitions and measurements across study participants are used for the same predictors, then differences in their associations with the outcome can be expected. For example, active lower digestive tract bleeding may be included as a possible predictor in a diagnostic model developed to detect colorectal cancer. This predictor 'blood in faeces' could be assessed in some study participants based on visible blood in the stool and in other participants using faecal occult blood testing. However, if these methods with different minimum detection levels are used interchangeably as a single predictor, 'blood in faeces' has the potential to introduce bias, especially if the choice of measurement method was based on prior tests or symptoms.

The potential for this bias is higher for predictors that involve subjective judgement, such as imaging test results. Here there is a risk of studying the predictive ability of the observer rather than that of the predictors.(1, 112-115) Where special skill or training is required, it may also be important to specify who assessed the predictor, for example, experienced consultant versus inexperienced trainee.

Example:
Perel et al. assessed the following predictors, all of which were recorded on the entry form for the CRASH-2 randomised trial: demographic characteristics (age and sex), characteristics of the injury (type of injury and time since injury), and physiological variables (Glasgow coma score, systolic blood pressure, heart rate, respiratory rate, central capillary refill time).(89) As the data used for the development of the prediction model came from a sub-study of a randomised trial and predictors were taken from the study entry form, it is likely – although not specifically described in the paper - that all predictors were defined and assessed in the same way for all participants. This signalling question would therefore be rated as PY. If data were derived from multiple sources such as in routine care data registries, where it is likely that different versions of the Glasgow coma scale were used or different definitions of injury type were used, then this signalling question would be answered as PN.

2.2 Were predictor assessments made without knowledge of outcome data?

Risk of bias is low when predictor assessments are made without knowledge of the outcome status often referred to as "blinding" or "masking". Blinding predictor assessment to outcome data is particularly important for predictors that involve subjective interpretation or judgment, such as predictors based on imaging, histology, history or physical examination. Lack of blinding increases the risk of incorporating the outcome information into the predictor assessments which likely increases their association leading to biased, inflated estimates of model performance.(1, 112-120)

Blinding predictor assessors to outcome information occurs naturally in prognostic studies using a prospective cohort design when prognostic predictors are assessed before the outcome occurs. This bias is more likely in studies using retrospective reporting of predictors (vulnerable to recall bias) or cross-sectional studies, such as diagnostic model studies, where predictors and outcomes are assessed within a similar time frame.(1, 112-121)

Most prediction model studies do not report information on blinding of predictors to outcome data.(122, 123) In prognostic studies, this signalling question should then be rated as NI (Table 8). However, the domain can still be rated as low risk of bias in the overall risk of bias assessment, because if predictors were measured and reported a long time before the outcome occurred it can be inferred as 'blinded to the outcome'. Note that even in prognostic studies predictors may sometimes still be assessed retrospectively after the outcome information has been collected, for instance predictors collected from re-interpretation of stored imaging information or when using a retrospective follow-up design. An example is the re-use of frozen tissue or tumour samples to measure novel predictors (biomarkers); such samples will already be linked to participant follow-up information, and thus measurement of the novel predictors may happen after the outcome has occurred and may not be blinded to outcome information.

Example:
Oudega et al. stated that "after informed consent was obtained, the primary care physician systematically documented information on the patient's history and physical examination by using a standard form on which the items and possible answers were specified. Patient history included sex, presence of previous DVT, family history of DVT, history of cancer (active cancer in the past 6 months), immobilization for more than 3 days, recent surgery (within the past 4 weeks), and duration of the 3 main symptoms (a painful, red, or swollen leg). Physical examination included the presence of tenderness along the deep venous system, distention of collateral superficial veins, pitting edema, swelling of the affected limb, and a difference between the circumference of the 2 calves (…) After history taking and physical examination, all patients were referred to the hospital for D-dimer testing and leg ultrasonography".(86)
Since it was reported that all participants had their history and clinical information, i.e. the predictors, collected prior to the D-dimer testing and were therefore also blind to the outcome, this signalling question should be answered as Y.

2.3 Are all predictors available at the time the model is intended to be used?

For a prediction model to be usable in a real-world setting, all predictors included in that model need to be available at the point in time where the model is intended to be applied, i.e. at the moment of prediction (Table 2). This sounds so straightforward that it should always happen. Unfortunately, some models include predictors or predictor information that could not be known at the time when the model would be used.

For example, when developing a prognostic model to be used *pre-operatively* to predict the risk of nausea and vomiting within 24 hours after surgery, the model should not include predictors such as *intra-operative* medication, unless this medication is pre-set and unchanged during surgery. Inappropriate inclusion of predictors not available at the time when the model would be used makes a model unusable and also inflates apparent model performance, by inclusion of predictors measured closer in time to the outcome assessment which are likely to be more strongly associated with the outcome. For predictors that are stable over time (e.g. gender and genetic factors), these aspects are not an issue.

In studies that aim to externally validate an existing prediction model, the study has high risk of bias when the model is validated while not having the data of each of the predictors (in that model) but validation is done anyhow using the model simply omitting these missing predictors. This is a common flaw in validation studies and effectively produces validation results for another model, rather than a validation of the intended original developed model. In these situations, this signalling question should be answered as N.

Example:
Rietveld 2004 aimed to develop and validate a prediction model for the diagnosis of a bacterial origin of acute conjunctivitis in children presenting in primary care with symptoms of this disease to decide on the administration of antibiotics.(90) All predictors should be available to the general practitioner during the initial consultation. The predictors in this study were indeed all obtained during history taking and the physical exam. The study should therefore be answered as Y for this signalling question. If the study had included laboratory testing (e.g. microscopy) amongst the predictors assessed, then this signalling question would be likely to be answered as N. This is due to the time delay involved in obtaining microscopy results, making it unlikely that the GP would have the results available during the initial consultation.

Rating the risk of bias for domain 2

Table 8 shows how the signalling questions should be answered and an overall judgement for domain 2 reached.

Applicability

Common reasons for concerns for the applicability in this domain are that definition, assessment or timing of predictors are not consistent with the review question. Predictors should be measured using methods potentially applicable to the daily setting (Table 5) that is addressed by the review. Primary studies that used specialised measurement techniques for predictors may yield optimistic predictions for the targeted setting of the review. For example, if a model should be used in a health setting with limited access to imaging, a study that developed a model including results of positron emission tomography (PET) might not be applicable, and so may be rated as high concern.

As for domain 1, there can be a subtle distinction between risk of bias and applicability assessment in this domain. Consider the example of active lower digestive tract bleeding as a predictor for colorectal cancer presence considered in signalling question 2.1. Such bleeding could be assessed based on visible blood in the stool or using faecal occult blood testing. Reviewers might focus their review to include diagnostic models that used only the 'visible assessment' as a predictor of colorectal cancer. With a systematic review focus on using a 'visible assessment' test, a primary study using a faecal occult blood test would raise applicability concerns.

Similarly, as for domain 1, in reviews that aim to estimate the average predictive performance of a specific model, heterogeneity in the observed performance of that model across the development study and validation studies is expected due to differences in definition and measurement of the predictors. If different definitions or assessment methods are used, some validation studies might find different predictive performance than others and should be judged as a concern for applicability. Sometimes researchers intentionally applied different definitions or measurement methods of predictors, for example using point of care rather than laboratory testing methods for certain blood values. Again, this might not be a problem if the explicit aim of the systematic review was to include all validations of a certain model, regardless of the definition and measurement method of the predictors in that model.

Domain 3: Outcome

This domain covers potential sources of bias and applicability concerns related to the definition and determination of the outcome. The ideal outcome determination would classify the outcome without error in all study participants.

In *diagnostic* model studies, the outcome is presence or absence of the target condition. Outcome determination, or verification, is measured using a reference standard (Box 2). For *prognostic* model studies, the predicted outcomes occur in the future, after the moment of prediction. For both diagnostic and prognostic models, the reference standard or outcome determination method may include a single test or procedure, a combination of tests (composite outcome), or a consensus by experts, e.g. an outcome adjudication committee.

The support for judgement box enables reviewers to describe how the outcome was defined, determined and in what time interval, and the information available when determining the outcome.

Risk of bias

There are six signalling questions to facilitate a risk of bias judgment for this domain (Table 9).

3.1 Was the outcome determined appropriately?

The rationale for this signalling question is to detect potential for bias due to outcome misclassification because suboptimal or inferior methods were used to determine the outcome. Errors in outcome classification can lead to biased regression coefficients, biased estimates of the intercept (logistic regression and parametric survival models models) or baseline hazard (Cox regression model), and thus biased performance measures of the prediction model.

When prediction model studies use data from routine care registries or from existing studies originally designed and conducted to answer a different research question, a careful appraisal is needed to determine appropriateness of methods used for determining the outcomes, sometimes using details from earlier publications about that study. In routine care registries, outcome data might not be recorded at all, or used methods may have been suboptimal and have missed or misclassified the outcome. In diagnostic studies, problems and bias due to misclassification of the target condition by suboptimal reference standard methods have been extensively studied.(113, 117, 124-128)

Similar to measurement of predictors (signalling question 2.1), the potential for bias is higher for outcomes that involve subjective judgement, such as imaging, surgical or even pathology procedures. Where special skill or training is required, it may also be important to specify who determined the outcome, for example, experienced consultant versus inexperienced trainee.

> **Example:**
> In Han 2014, "there were two defined outcomes for each of the models: one was mortality at 14 days, and the other was unfavourable outcome at 6 months", defined by the authors based on the Glasgow Outcome Scale (GOS) as "severe disability, vegetative state, or death". As the outcomes, mortality and the three categories based on the definition of GOS, use well established, appropriate measures for outcome determination, the signalling question should be answered as Y.
> Problems could arise if the Glasgow Outcome Scale had been measured by assessors who are not trained in determining this outcome. Despite the limited number of categories, misclassification is not uncommon for the GOS.(129, 130) The use of inexperienced assessors could lead to a less appropriate (PN or NI) answer for this signalling question.

3.2 Was a pre-specified or standard outcome definition used?

This signalling question aims to detect the potential risk of bias where model performance has been inflated by selecting an outcome definition that produces more favourable results.(131)

The risk of bias is low when a pre-specified or standard outcome definition is used, substantiated by a definition from clinical guidelines, previously published studies or a published study protocol. Risk of bias is higher if an atypical threshold on a continuous scale has been used for defining an "outcome being present". Biased model performance can occur if authors test multiple thresholds to obtain the most favourable outcome definition to achieve the best estimate of model performance. For example, a biased assessment of model performance would result if authors used a continuous scale such as the Glasgow Outcome Scale (GOS) ranging from 3 to 15 and chose a threshold for classifying "good" and "poor" outcomes based on achieving the best model predictive performance.

Composite outcomes can also introduce risk of bias. For example, authors may introduce bias by adjusting a composite outcome definition to favour better model performance by leaving out typical components or including non-typical events.

For many outcomes, there is consensus on outcome definitions, including thresholds and preferred composite outcome definitions. The COMET initiative (Core Outcome Measures in Effectiveness Trials,

http://www.comet-initiative.org) was set up to facilitate development of agreed standardised sets of outcomes. Determining whether standard or non-standard definitions have been used may require specialist clinical knowledge.

> **Example:**
> In Han 2014, "there were two defined outcomes for each of the models: one was mortality at 14 days, and the other was unfavourable outcome at 6 months, defined by the authors based on the Glasgow Outcome Scale (GOS) as severe disability, vegetative state, or death". Given that both, mortality and the three categories based on the definition of GOS, are well established outcomes, i.e. standard outcome definitions were used, the signalling question should be answered as Y.
> If the authors instead of using a standard definition had amended the categories of the GOS based on their own clinical experience or following internal hospital guidance, clinical judgement should be used to decide whether these changes still constitute a standard outcome determination or whether the signalling question should be answered as PN or N.

3.3 Were predictors excluded from the outcome definition?

Outcomes should ideally be determined without information about the predictors (see signalling question 3.5), but in some cases it is not possible to avoid including predictors, for example when outcomes require determination by a consensus panel using as much information as is available. If a predictor in the model forms part of the definition or assessment of the outcome that the model predicts, it is likely that the association between the predictor and outcome will be overestimated, and estimates of model performance are optimistic; in diagnostic research this problem is generally referred to as incorporation bias.(105, 112, 116, 118, 120, 132-135)

Where outcomes are difficult to determine by a single procedure (e.g. a single reference test), determination of an outcome presence or absence may be based on multiple components or tests (as in the World Health Organisation criteria for the diagnosis of myocardial infarction) or even on all available information including the predictors under study. The latter approach is known as consensus or expert panel outcome measurement and also susceptible to incorporation bias.(136)

> **Example:**
> Aslibekyan 2011 aimed to develop a cardiovascular risk score based on the ability of predictors such as dietary components, physical activity, smoking status, alcohol consumption, socioeconomic status and measures of overweight and obesity to predict non-fatal MI.(87) The study reported that MI was defined according to World Health Organization criteria. These criteria include cardiac biomarkers, electrocardiogram, imaging, or autopsy confirmation. Since the lifestyle and socioeconomic predictors used for modelling in Aslibekyan 2011 do not form any part of this definition of MI, the study would be rated as Y for this signalling question.
> If the study had included a cardiac biomarker (e.g. troponin T at initial hospital presentation) amongst the predictors assessed, then this signalling question would be likely to be rated as N. This is because the initial troponin T measurement may have formed part of the information used to determine the outcome (MI).

3.4 Was the outcome defined and determined in a similar way for all participants?

The outcome should be defined and determined in the same way for all study participants, similar to predictors (signalling question 2.1).

Outcome definition and measurement should include the same thresholds and categories to define the presence of the outcome across participants. Where a composite outcome measure is used, the results of individual components should always be combined in the same way to establish the outcome presence or absence. When using a consensus or panel-based outcome committee, the same method for establishing the outcome, for example majority vote, should be used.(132, 136, 137)

Risk of bias can arise when participants differ in the way their outcomes are determined, for example due to variation in methods between research sites in a multi-centre study. Risk of bias is also increased when prediction model studies are not based on pre-designed studies, but on data collected for a different purpose, such as routine care registry data, where inherently different outcome definitions and measurements are likely to be applied. Risk of bias is also higher when different measurement methods have different accuracy for determining the presence of an outcome (differential outcome verification) and the direction of bias is not easy to predict. For example, in a *prognostic* model study aimed at predicting the future occurrence of diabetes in healthy adults, the presence of diabetes in an individual can be determined in various ways which all may have different ability to determine diabetes presence or absence, e.g. using fasting glucose levels, oral glucose tolerance test or self-reported. The potential for bias is higher when outcomes require more subjective interpretation. Similarly, outcomes measured on multiple occasions such as clinic visits are at risk of bias, particularly if the frequency of measurement is different between participants; more measurement occasions increase the likelihood of detecting the outcome.

In *diagnostic* studies, researchers sometimes explicitly did not or could not apply the same outcome measurement in each individual. For instance, in cancer detection studies, pathology results are likely to be available as a reference standard only for those participants who have some positive result on a preceding index test such as an imaging test. Two situations may then occur: *partial verification*, when outcome data are completely missing for the subset of participants who tested negative on the index test and for whom there is no reference standard result, and *differential verification*, when participants who are not referred to the preferred reference standard are assessed using an alternative reference standard of differing, usually lower, accuracy.(107, 112, 118, 120, 132-135, 138) These differences in outcome determination affect the estimated associations of the predictors with the outcome and thus the predictive accuracy of the diagnostic models., methods to account for partial and differential verification have been described.(139-142)

> **Example:**
> Han et al. 2014 validated a model to predict "unfavourable outcome after six months" in patients with severe traumatic brain injury.(88) The outcome was determined using the Glasgow Outcome Scale (GOS; levels 1 to 3 on the 5-point GOS) for all patients included in this single centre study. This should be answered as Y.
> If a hospital in the study had used a different instrument to measure the outcome of interest, e.g. the Functional Status Examination (FSE) rather than the GOS, this would constitute a potential risk of bias as these tools are not directly comparable. Then this signalling question would be answered as PN or even N to highlight the potential risk of bias.

3.5 Was the outcome determined without knowledge of predictor information?

The outcome is ideally determined without knowledge of information about the predictors. This is comparable to intervention trials where the outcome is ideally determined without knowledge of the treatment assignment. Knowing predictor results may influence outcome determination, and could lead to biased predictive accuracy of the model, usually due to overestimation of the association between predictors and outcome.(112, 116, 118, 120, 133-135) This risk is lower for objective outcomes, such as death from any cause or whether a child birth was natural or by caesarean section, but higher for outcome determinations requiring interpretation, such as death from a specific cause.

Some outcomes are inherently difficult to determine using a single measurement method or test. As discussed in signalling question 3.3, sometimes diagnostic and prognostic research cannot avoid the use of a consensus panel or end-point committees, where outcome determination includes knowledge of predictor information. If the explicit aim is to assess the incremental value of a particular predictor

or when comparing the performance of competing models (e.g. when validating multiple models on the same data set), the importance of blinded outcome determination increases to prevent overestimation of the incremental value of a particular predictor, or to prevent biased preference for one model to another.

Review authors should carefully assess whether predictor information was available to those determining the outcome. If predictor information is present when determining the outcome or when it is unclear, the potential consequences should be judged in the overall judgment of bias of this domain. This overall judgment should be made taking into account the subjectivity of the outcome of interest and the underlying review question.

> **Example:**
> In the diagnostic prediction model study of Rietveld et al., the outcome of interest was a bacterial infection of the eye established by culture as the reference standard procedure.(90) Reading of the results of the cultures was somewhat subjective. Therefore, the authors of the paper explicitly inform the reader about the degree of blinding in their study: "The general practitioners did not receive the culture results, and the microbiologist who analysed the cultures had no knowledge of the results of the index tests" [read: the candidate predictors of the study]. The signalling question "Was the outcome determined without knowledge of predictor information?" should therefore be answered as Y.

3.6 Was the time interval between predictor assessment and outcome determination appropriate?

This signalling question is to detect situations where the time interval between predictor assessment and outcome determination is inappropriate, either too long or too short. Such judgement requires clinical knowledge to determine what an appropriate time interval is, and also depends on the clinical context.

In *diagnostic* studies where the model is predicting whether the outcome (i.e. target disease determined by a reference standard) is present at the moment of prediction (Box 2), ideally the assessment of predictors (index tests) and outcome should occur at the same point in time. In practice, there may be a time interval between the moment of assessing the predictors and outcome where the diagnostic outcome classification could change, either improving or worsening. Sometimes determining the outcome presence requires clinical follow up over a time period, so a delay between predictor and outcome assessment is built into the study design, as a critical feature to reduce bias (see the example study of Oudega et al).

A delay between predictor assessment and outcome determination of a few days may not be problematic for chronic conditions, while for acute infectious diseases even a short delay may be problematic. Conversely, when the reference standard involves follow-up, a minimum length may be required to capture the increase in symptoms or signs indicating that the disease was present at the moment when the predictors were assessed. Sometimes biological samples for predictor assessment and outcome determination are taken at the same time point, so the time interval during which the disease status could change is effectively zero even if the reference standard procedure on the sample is completed at a later time point.

In *prognostic* studies, the time interval between the moment of assessing the predictors and outcome determination may also have been too short or too long to capture the clinical relevant outcome of interest.

For both *diagnostic and prognostic* models, there are two ways bias can present. Firstly, bias can result if outcomes are determined too early when relevant outcomes cannot be detected or the number of outcomes is unrepresentative. For example, in a model diagnosing the presence of metastases at the time of surgical removal of colorectal cancer tumour, the detection of metastases can be biased by the time point of follow-up used for the reference standard. Choice of a time point that is too early can introduce bias in the number of metastases detected, as due to limitations in current detection methods; at earlier follow-up times metastases may not have grown to a large enough size for detection. Secondly, the type of outcome may also be different depending on the time interval. For example, the metastases detected at earlier times might be mainly liver metastases, whereas at one year follow-up more bone metastases may be detected. A risk of bias then occurs if the length of interval between predictor assessments and outcome determination results in either determination of a potentially unrepresentative number of outcomes or type of outcomes (i.e. metastatic locations).

The aim of a review may be specifically in either the short and long-term prognosis of a certain condition, so the time interval between predictor assessment and outcome determination is also relevant to the applicability of a study to the review question.

> **Example:**
> In Rietveld et al. where a diagnostic model is developed to predict bacterial cause in conjunctivitis eye infection, risk of bias in the time interval is minimised as the same clinic visit is used to measure predictors from patient questionnaires and physical examination, and to collect conjunctival samples for determination of the outcome of bacterial infection.(90) Although the reference standard results require culture for more than 48 hours, this is not relevant to bias, as culture results reflect disease at the time of sample collection. This signalling question would be answered as Y indicating a low potential for bias.
> In Aslibekyan et al. where a model is developed to predict myocardial infarction, this signalling question should be answered NI due to lack of information on the time interval between predictor measurement and the outcome determination for myocardial infarction.(87) Different time intervals could alter the number of myocardial infarction events that would be detected.

Rating the risk of bias for domain 3

Table 9 shows how the signalling questions should be answered and an overall judgement for domain 3 should be reached.

Applicability

The applicability question for this domain considers the extent to which the outcome predicted in the developed or validated model matches the review question. If different definitions, timing or determination methods are used, this should be judged a concern for applicability. For example, the study might use a composite outcome which consists of components different to the ones included in the outcome definition of the review question.(143)

In reviews that aim to estimate the average performance of a specific model across the included validation studies, heterogeneity in performance between the validation studies is expected due to differences in definition and measurement of the outcome. Sometimes researchers intentionally applied different outcome definitions or measurement methods. This might not be a problem if it was the explicit aim of the systematic review to include all validations of the model, regardless of outcome definition and measurement method.

Domain 4: Analysis

The use of inappropriate analysis methods, or the omission of important statistical considerations, increases the potential for bias in the estimated predictive performance of a model. Domain 4

examines whether key statistical considerations were correctly addressed. Some of these aspects require specialist knowledge and we recommend that this domain is assessed by at least one individual with statistical expertise in prediction model studies. The support for judgement box should list and describe the important aspects needed to address this domain.

Risk of bias

There are nine signalling questions to facilitate a risk of bias judgment for this domain (Table 10).

4.1 Were there a reasonable number of participants with the outcome?

As applies for all medical research, the larger the sample size the better, as it leads to more precise results, i.e. smaller standard errors and narrower confidence intervals. For prediction model studies, it is not just the overall sample size that matters but more importantly the number of participants with the outcome. For a binary outcome, the effective sample size is the smaller of the two outcome frequencies, 'with the outcome' or 'without the outcome'. For time-to-event outcome, the key driver is the total number of participants with the event by the main time-point of interest for prediction. More importantly, in prediction model studies the number of participants with the outcome not only influences the precision but also affects predictive performance, i.e. is a potential source of bias. What is considered a reasonable number of participants with the outcome (yielding low risk of bias) differs between model development and validation studies.

Model development studies

The performance of any prediction model is to varying extents overestimated when the model is both developed and its performance assessed on the same dataset.(49, 81, 147, 148) This overestimation is larger with smaller sample sizes and notably with smaller number of participants with the outcome. Concerns about optimistic performance are exacerbated when the predictors included in the final model are selected from a large number of candidate predictors, relative to a low number of participants with the outcome, and when predictor selection was based on univariable analysis (see signalling question 4.5). Sample size considerations for model development studies have, historically, been based on the number of events-per-variable. More exactly, it is the number of events relative to the number of regression coefficients that need to be estimated for the candidate predictors. For example, a candidate predictor with six categories will require five degrees of freedom (five regression coefficients are estimated). Also, the word *candidate* is important as it is not the number of predictors included in the final model but rather the total number of predictors that were considered during any stage of the prediction model process.

While an EPV of at least 10 has been widely adopted as a criterion to minimize overfitting(149-151), recent studies have shown that EPV of 10 has no scientific basis(146) and various authors suggested higher EPVs of at least 20.(146, 152, 153). In general, studies with fewer than 10 EPV are likely to suffer from overfitting, whilst those with an EPV of more than 20 are less likely to suffer from overfitting. However, the sample size needed to minimize overfitting is context specific, dependent on outcome prevalence, overall model performance (R-squared), and the predictor distributions.(144-146) Therefore it may be difficult to decide whether an appropriate sample size was used, especially when EPV is between 10 and 20. Prediction models developed using machine learning techniques often require substantially higher EPV to minimize overfitting, with an EPV of at least 200 often needed.(57)

Hence, the smaller the effective sample size and the lower the EPV, the higher the risk the final prediction model has included spurious predictors (so-called overfitted models) or failed to include important predictors (underfitting). Overfitting and underfitting are likely to yield biased estimates of the model apparent predictive performance.(49, 50, 81, 147, 148, 154) With small EPV, authors need

to quantify the extent of misfitting of the developed prediction model, for example by using internal validation techniques. Based on this internal validation, optimism-adjusted estimates of model performance can be produced and model parameters adjusted (i.e. shrink regression coefficients) to decrease this bias (see signalling question 4.8).

Model validation studies

In a validation study, the aim is to quantify the predictive performance of an existing model using a separate dataset from the model development.(8, 49, 81, 155-157) Emphasis in a validation study is on accurate and precise estimation of model performance so that meaningful conclusions can be drawn. Sample size recommendations for validation studies are that at least 100 participants with the outcome are needed, otherwise the risk of biased estimates of model performance increases.(77, 78, 158)

> **Example:**
> Aslibekyan et al. developed two prognostic models (one including only easy to obtain predictors and one extended with various dietary and blood markers) to predict the risk of developing myocardial infarction (MI).(87) Although the authors used a case-control study design and many inclusion and exclusion criteria, they ended up with 839 cases with an MI for developing score 1 and 696 for score 2. The exact number of candidate predictors is not explicitly mentioned but from the methods and supplementary tables 1 and 2 we can estimate that the authors likely used 20 to 30 predictors or rather degrees of freedom as they categorised several continuous predictors into quintiles. This indicates that the EPV is between (taking the smallest number of events) 696/20 (i.e. 35) and 696/30 (i.e. 23). As the EPV in either case is much larger than 10, this signalling question should be answered Y, indicating a low risk of bias.
> Oudega et al. validated a diagnostic model for detecting the presence of deep vein thrombosis (DVT) in patients who consulted with their primary care physician about symptoms suggestive of DVT.(86) The total sample size of their validation study was 1295 patients with symptoms of DVT of whom 289 had an DVT (as detected by D-dimer and leg ultrasonography). Since, the number of events is larger than the recommended 100 events needed for validation, the signalling question, for this validation study, should be answered as Y, indicating a low risk of bias. If this number was lower, e.g. 80 or 40 patients with DVT, then the answers for this example would be PN or N, respectively.

4.2 Were continuous and categorical predictors handled appropriately?

Dichotomisation of continuous predictors, such as age and blood pressure, should be avoided.(159-161) Dichotomisation requires choosing an often arbitrary cut-point value, for example above which participants, are classified as high (or abnormal) and below which they are classified as low (or normal). The usual fallacious argument for the approach is to aid clinical interpretation and maintain simplicity. However, it leads to loss of information and reduced predictive ability of a prediction model including dichotomised continuous predictors can be substantial.(159-162)

For example, dichotomising a variable at the median value has been shown to reduce power by about the same amount as discarding a third of the data.(163) Also, the range of model predicted risks across the spectrum of predictor values is lost: individuals just below the cut-point are assumed to have a different risk from those just above the cut-point, even though their predictor values barely differ. Conversely, two individuals with very different values but both above (or both below) the cut-point are assumed to have identical risks. Linear (or non-linear) relationships between the predictor and outcome risk are therefore lost. When a predictor is categorised using widely accepted cut-points, although information has been lost, there is a low risk of bias since the predictor cut-point was pre-defined.

Model development studies

A developed model is at low risk of bias when included predictors are kept as continuous. The association between the predictor and outcome risk should still be examined as linear or non-linear by using, for example, restricted cubic splines or fractional polynomials.(49, 81, 164)

A developed model is at high risk of bias when dichotomised continuous predictors are included, especially when (i) cut-points were chosen via data-dredging on the same data set, for example to identify the 'optimal' cut-points that maximises predictor effects or minimises associated p-values;(159-162) and (ii) a selection procedure was used to identify the 'significant thresholds'.(49, 81)

Risk of bias is decreased when the model uses categorisation of continuous predictors into four or more groups, rather than dichotomising, especially when it is based on widely accepted cut-points.(160, 162) However, for classification of low-risk of bias, it should be clear that the number and placement of cut-points of predictors was chosen in advance of data analysis. For similar reasons as discussed for signalling question 4.1, an internal validation followed by optimism-adjustment of model performance and prediction model parameters, also decreases the risk of bias (see also signalling question 4.8). For model development studies which have dichotomised continuous predictors after the data analysis and did not adjust for it by applying internal validation and shrinkage techniques, this signalling question should be answered as N.

Model validation studies

In model validation studies, the model as originally fitted in the development data should be evaluated on its predictive accuracy in the validation dataset. This means that the originally reported intercept (or baseline hazards) and regression coefficients are used for exactly the same format of the predictors. For example, if body mass index (BMI) is originally included as dichotomised in the model, then validation studies should use BMI values dichotomised at the same cut-point and not BMI as continuous or dichotomised using a different cut-point. If predictors do not have the same format in the validation as used in the development model, the validation might be considered at high risk of bias since the predictor-outcome association (the regression coefficient) of BMI from the development study was effectively used in the validation study for a different version of the predictor.

> **Example:**
> Oudega et al. validated the Wells rule for identifying individuals with deep vein thrombosis (DVT).(86) However, the authors comment that "the last item of the rule—presence of an alternative diagnosis— has never been unambiguously defined and often causes controversy among users of the rule. In our study, physicians were asked to give their own assessment of the patient's probability of having DVT by using a score of 1 to indicate high probability of DVT, no alternative diagnosis likely; 2 to indicate moderate probability of DVT, alternative diagnosis possible; or 3 to indicate low probability of DVT, alternative diagnosis certain. To tailor the judgment of the physician on this item, 7 common alternative diagnoses for patients with suspected DVT were provided on the study form. If a low or moderate probability was assigned to a patient, we subtracted 2 points from the Wells score in the analysis". Since this is not a true deviation from the original definitions, this signalling question should be answered as Y.
> Perel et al. developed a prediction model (CRASH-2) for early death in patients with traumatic brain injury, and during model development they take a three category variable 'type of injury' (penetrating, blunt, or blunt and penetrating) and analyse it as a two category variable (penetrating versus a combined category of blunt and penetrating), the rationale for this is not given.(89) Nevertheless, continuous variables were analysed as continuous in the model development, and so the collapse from 3 to 2 categories for this variable was probably due to few participants or events being in the 'blunt' category. Further, the type of injury was not subsequently included in the final model, and so it is unlikely that reduction in predictor categories was done in order to improve statistical significance for this predictor. Therefore, we would rate the signalling question as Y. When externally validating the CRASH-2 model, the authors "applied the coefficients of the model developed in

CRASH-2" and appear to use the same predictors and scale as originally coded, and thus an answer of Y seem appropriate.

4.3 Were all enrolled participants included in the analysis?

As applies to all types of medical studies, all participants enrolled into a study should be included in the data analysis, otherwise there is a potential for risk of bias.(46, 112, 165, 166) This signalling question relates to exclusion of participants from the original study sample who met the inclusion criteria. It is not about inappropriate inclusion criteria (which are addressed in signalling question 1.1) and not about the *handling* of missing data in predictors or outcomes (which is covered in signalling question 4.4).

Enrolled participants are often excluded due to uninterpretable (unclear) findings, outliers or missing data in predictors or outcomes (due to loss to follow up). Outlier, uninterpretable or missing values occur in all types of medical research. Omitting enrolled participants from analysis can lead to biased predictor-outcome associations and biased predictive performance of the model, if the remaining analysed individuals are not a completely random but rather a selective subsample. The relationship between predictors and outcomes is then different for the analysed versus the excluded participants. For example, excluding participants from the study sample where predictor values (e.g. imaging or lab test results) were unclear likely yields a study sample with participants in the extremes of the predictor range. This in turn may result in biased, overestimated, model discrimination.(166) When only a low percentage of enrolled participants are not included in the analysis, there may only be a low risk of bias. However, a minimal or acceptable percentage is hard to define as it depends on which participants were excluded, and whether it was a selected subsample or not. The risk of bias increases with an increasing percentage of participants excluded.

Prediction model studies based on routine care databases or registries, where participants are not formally enrolled in some study and data are originally collected for other reasons, are particularly susceptible to this form of bias. When such data sources are used for model development or validation, participant selection should be based on clear inclusion criteria. We note that in such routine care datasets, the extent of potential bias may sometimes be unclear due to unreported information relating to specific inclusion criteria and reasons for exclusion of included participants.

Example:
In Han et al., all 300 participants met the inclusion criteria for validation of three versions of the IMPACT models for TBI referred to as core, extended and laboratory IMPACT models.(88) Thirty-six participants (12%) were excluded from validation of the laboratory version of the IMPACT model due to missing data on blood glucose level, however all participants could be included for both the core and extended IMPACT models. For assessment of the core and extended CT models, the signalling question would be answered as Y as all participants are included in the analysis. For the assessment of the laboratory model, the signalling question would be answered as either PN or PY, depending on the concern from exclusion of 36 (12%) of participants from the analysis. This would depend on clinical knowledge and judgement of whether the missing glucose measurements are likely to be associated with the severity of patient TBI.

4.4 Were participants with missing data handled appropriately?

As noted in the previous item, simply excluding enrolled study participants with any missing data from the analysis leads to biased predictor-outcome associations and biased model performance when the analysed individuals are not a completely random sample from the original full study sample but rather a selective subsample.(167-177) When there is no mention of missing data in a study report, it is likely that participants with any missing data have simply been omitted from any analyses (so-called available case or complete-case analysis) as statistical packages automatically exclude individuals with

any missing value on any of the data analysed unless prompted to handle otherwise. Reviews showed that available or complete case analysis is the most common way to handle missing data in prediction model studies.(68, 178-186)

The most appropriate method for handling missing data is multiple imputation as it leads to the least biased results with correct standard errors and p-values.(167-173, 175-177) In prediction model studies multiple imputation is superior in terms of bias and precision to other methods, both in model development(173, 176, 187) and validation studies(176, 188-190). In contrast to uninterpretable or outlier data, the use of a separate category to capture missing data is not an appropriate method for handling participants with missing data. The use of this missing indicator method leads to biased results in prediction model studies and this signalling question should then be rated as N.(172, 177) The risk of bias due to missing data increases with increasing percentages of missing data, but a minimal acceptable percentage which can be used as a threshold for a low risk of bias is hard to define.(173) To judge a possible risk of bias, it is useful when authors provide the following: the distributions (percentage, mean or medians) of the predictors and outcomes between both groups (excluded versus analysed participants); or a comparison of the predictor-outcome associations and the model predictive performance with and without inclusion of the participants with missing values. If results are similar with and without participants with missing values, there is a strong indication that the results of the analysis are less likely to be biased. If such comparison is not presented and investigators have not used an imputation method, we recommend to rate this signalling question as PN or N, certainly if a relevant proportion of participants are excluded due to missing data.

Sometimes, when a model is validated in other data and a predictor of the model is systematically missing (e.g. not measured), authors validate the model by simply omitting the predictor from the model and validate the original model (i.e. the original predictor weights or regression coefficients) without that predictor. This leads to a high risk of bias and such studies should be rated as N for this question. If the model had originally been fit without the omitted predictor, all the remaining predictor coefficients would be different.

Example:
Perel et al. developed a prognostic model from a data set with 'very low amount of missing data and therefore they did a complete case analysis'.(89) The authors showed in the same paper an external validation of this developed model where they applied multiple imputation. It was neither clear from the development study how low the number of participants with missing data was nor was any comparison given between the completely observed and excluded set of participants, making it hard to judge whether there was some risk of bias in the model development. In the validation study the authors used multiple imputation indicating that they know the procedure; if it was needed to multiply impute missing data in the development sample, they likely would have used multiple imputation as well. Accordingly, this signalling question should strictly be answered as NI for the development and Y for the validation part of the paper, although PY for the development part would also be possible.
In Aslibekyan et al., the authors state that for their model development complete case analysis, with 10% of participants being excluded, was used. No information was provided to confirm that complete case analysis was a valid approach, i.e. that the included and excluded participants were similar, or that the included participants approximated to a completely random subset of the original study sample.(87) Accordingly this signalling question should be rated N for the development part. For the model validation, there was no mention of missing data or handling of missing data. Accordingly, the answer for this signalling question for the model validation should strictly be NI, but perhaps even PN as all clinical studies tend to have some missing data.

4.5 Was selection of predictors based on univariable analysis avoided? (Model development studies only)

Often many features are available in a dataset that could be used as candidate predictors, and in many studies researchers want to reduce the number of predictors during model development to produce a simpler model.

In a univariable analysis, individual predictors are tested for their association with the outcome. Often researchers select the predictors with a statistically significant univariable association (e.g. at p-value < 0.05) for inclusion in the development of a final prediction model. This method can lead to incorrect predictor selection for developing the model as predictors are selected based on their statistical significance as a single predictor rather than in their context with other predictors.(49, 81, 191) Bias occurs when univariable modelling results in omission of variables from the model because some predictors are only important after adjustment for other predictors, known from previous research to be important, did not reach statistical significance in the particular development set, for example due to small sample size. Also, predictors may be selected in univariable selection based on spurious (accidental) association with the outcome in the development set.

A better approach to decide on omitting, combining or including the candidate predictors in the multivariable modelling is to use non-statistical methods, i.e. without any statistical univariable pre-testing of the associations of the predictors with the outcome. Better methods include those based on existing knowledge of yet established predictors in combination with the reliability, consistency, applicability, availability and costs of predictor measurement relevant to the targeted setting. It is recommended that predictors with clinical credibility and those already well established are included and retained in a prediction model regardless of any statistical significance.(49, 81, 192) Alternatively, some statistical methods that are not based on prior statistical tests between the predictor and the outcome, can be used to reduce the number of modelled predictors, for example principal components analysis (PCA).

During modelling, predictor selection strategies may be used to omit predictors (e.g. backwards selection procedures) and to fit a smaller, simpler final model.(49, 81, 192) However, the effects of using such multivariable predictor selection strategies on the potential overfitting of the prediction

model to the development data at hand should be tested using internal validation and optimism-adjustment strategies which are discussed in signalling question 4.8.

When the model development correctly avoids univariable selection or there is no evidence of univariable selection for predictions prior to the multivariable modelling, studies should be rated as Y or PY. When predictors are selected based on univariable analysis prior to multivariable modelling, the signalling question for these studies should be answered as N.

> **Example:**
> In Perel et al., before developing the model, potential users of the model were consulted to identify candidate predictors and interactions based on known importance and convenience to the clinical settings of pre-hospital, battlefield and emergency departments.(89) The researchers then included all so defined candidate predictors in the multivariable analysis. Decisions on which predictors were eventually retained in the final prediction model were based on clinical reasoning, availability of predictor measurement at the time the model would be used, and practicalities of collecting predictors using equipment in the clinical settings. Although there is a possibility that other predictors could have been considered important, the choice of predictors was not based on potentially biased univariable selection of predictors. The study would therefore be answered as Y for this signalling question.
> In Rietveld et al., predictor selection based on univariable analysis (p value of ≤ 0.10) was used to select predictors for the multivariable model.(90) This study would be therefore answered as N for this signalling question. If all predictors had been entered into multivariable analysis without the prior univariable selection, an answer of Y would have been given.

4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?

The development and validation of prediction models must ensure that the statistical methods used and their underlying assumptions are appropriate for the study design and type of outcome data analysed. Here, we draw attention to some key considerations related to complexities in the data that can lead to risk of bias of the estimated predictive performance of the model if not appropriately accounted for in the analyses.

As discussed under signalling question 1.1, if a case-cohort or a nested case-control design was used for a prediction model then the analysis method must account for the sampling fractions (from the original cohort) to allow for proper estimation of the absolute outcome probabilities.(98, 100, 106, 110) For example, in a diagnostic prediction model (development or validation) study that used a nested case-control design where a fraction of all the controls were sampled from the original cohort, a logistic regression in which the controls are weighted by the inverse of their sampling fraction needs to be applied instead of a standard logistic regression, otherwise the predicted risks by the model will be biased. When such appropriate adjustments for sampling fractions are made, they alleviate the risk of bias concerns raised in signalling question 1.1. If not done, one should score a N only once to either signalling question 1.1 or this signalling question.

For prognostic models to predict long term outcomes in which censoring occurs, it is important that a time-to-event analysis such as a Cox regression is used to include censored individuals up to the end of their participant follow-up. It is inappropriate to use logistic regression models that simply exclude censored participants with incomplete follow-up. Using a flawed logistic regression approach leads to a selected dataset with fewer individuals without the outcome which biases predicted risks as individuals with outcome are overrepresented. Time-to-event analysis correctly deals with these censored individuals.

When there are prominent competing risks these should also be accounted for in the time-to-event analysis when developing a prognostic model. An example of competing risks would be in a model for occurrence of a second hip replacement where death in elderly patients with a first hip replacement may occur before the second hip replacement. If competing risk is not correctly accounted for then absolute risk predictions will be overestimated and biased as patients with the competing event are simply censored.(193)

Also, correct modelling methods are needed where multiple events per individual can occur, such as in a model of epilepsy seizure, where some individuals experience more than two seizures. Multi-level or random effects (logistic or survival) modelling methods would be needed to avoid underestimation and bias in predictor effects.(194-197)

Statistical expertise will be required to identify these and potentially other issues in specific studies. The issues we have highlighted here will typically be the most important to be aware of in prediction modelling studies. If it is deemed that key statistical complexities are being ignored in a study, there may be a strong indication of a high risk of bias on this signalling question.

> **Example:**
> In Aslibekyan et al., a conditional logistic regression model was used to *develop* a prognostic prediction model for MI.(87) Included participants provided data between 1994 and 2004, however, it is unclear whether all individuals had predictor values recorded at the start of the period, or whether they could enter post-1994 and thus have a shorter follow-up. If all individuals entered with predictor values at 1994, then the model would predict risk of MI by 10 years (i.e. by 2004) and be interpretable. However, if some individuals entered after 1994, then the interpretation and bias of the logistic model is a concern because predictions are not specific to a particular time-period and the length of follow-up is being ignored. If participants had different times of follow up, it would be better for a survival analysis model to be fitted to allow risk predictions over time and delayed entry of participants. Further, it is not clear how prevalent the competing risk of death due to other non-MI conditions was, even though the included population went up to an age of 86 years. Such issues may be a consequence of the case-control (rather than cohort) nature of the study. Thus, risk of bias was not avoided (PN) due to these statistical complexities.
> In Rietveld et al., the development of a diagnostic model using standard logistic regression was relatively straightforward as the developed model aimed to predict risk of having a bacterial conjunctivitis using a full cohort approach (without sampling) and therefore did not involve follow-up, censoring or competing events.(90) In this case, the signalling question should be answered as Y.

4.7 Were relevant model performance measures evaluated appropriately?

Box 4 provides an overview of the various performance measures of a multivariable prediction model. PROBAST is designed to assess studies on multivariable models that are developed or validated to make predictions in individuals, i.e. *individualised predictions* (Box 1). Accordingly, to fully gauge the predictive performance of a model, both model calibration and discrimination (such as the c-index) addressing the entire range of the model predicted probabilities, need to be assessed.(7, 8) If calibration and discrimination are not assessed, the study is at risk of bias as the ability or performance of the model to provide accurate individual probabilities is not completely known (Box 4).

When calibration plots or tables are observed with small numbers of groups (e.g. possibly due to a small sample size with too few events), judgment of the plot is required to rate this signalling question properly. In the absence of a calibration plot or table comparing predicted versus observed outcome probabilities, studies reporting only a statistical test of calibration should be rated N for this signalling question.

Additionally, the methods used to assess model calibration and discrimination should also be

appropriate for the outcome the model is predicting. Approaches used to assess calibration and discrimination for models predicting a binary outcome developed using logistic regression will not be suitable for models predicting long term outcome occurrences, such as 5-year mortality or survival, using Cox regression as censoring needs to be accounted for. Failure to account for censoring when assessing prognostic model calibration and discrimination – either in a development or validation study - means the study should be answered as N or PN for this signalling question.

Some studies additionally provide classification measures such as sensitivity, specificity, predictive values or reclassification measures, such as the net reclassification index (NRI), to indicate a model predictive performance, sometimes without providing the model calibration and c-index (Box 4). Classification measures are most commonly provided in diagnostic model studies. Estimation of classification, as well as reclassification, parameters requires the introduction of one (or more) thresholds in the range of the model predicted probabilities. Using thresholds allows the reporting of model predictive performance at potentially clinically relevant probability thresholds, as opposed to entire range of the model predicted probabilities. Nevertheless, the use of probability thresholds typically leads to loss of information, since the entire range of predicted probabilities of the model is not fully utilised, and choice of thresholds can be data driven rather than pre-specified based on clinical grounds (see also signalling question 4.2). This practice can cause substantial bias in the estimated (re)classification measures, certainly when thresholds are chosen to maximise apparent performance.(84, 198) When the choice of threshold is not pre-specified, these methods are subject to risk of bias and this signalling question should be answered N. Also, when classification and reclassification measures are reported without model calibration, this signalling question should be answered as N. Before categorising model predicted probabilities, calibration is needed to understand whether the predicted probabilities are correct (Box 4).

> **Example:**
> In the study by Rietveld et al., the authors assessed the calibration by calculating the Hosmer-Lemeshow test, which resulted in a p-value of 0.117; this was interpreted that the model was well calibrated.(90) If this was the only measure to assess calibration of the model this signalling question would be rated as N as such p-value does neither indicate whether there was any miscalibration nor the magnitude of any miscalibration. However, in Table 4 the authors present the mean predicted probabilities with confidence intervals across subgroups and the corresponding observed outcome frequencies. This calibration table gives an indication of the model calibration, such that the answer to the signalling question for this study would be PY.
> In the validation of their model for predicting early death in patients with traumatic bleeding, Perel and colleagues evaluated calibration by presenting calibration plot of observed risks against predicted risks grouping by tenth of predicted risk.(89) Presenting calibration in this format allows the reader to judge the accuracy of the model over the entire probability range. The plot could be enhanced by overlaying the figure with a non-parametric (lowess) smoother. The authors also reported a c-index, enabling readers to judge the discrimination ability of the model although there was no 95% confidence interval to indicate the uncertainty of the estimate. This study would be at low risk of bias and answered as Y for this signalling question.

4.8 Was model overfitting and optimism in model performance accounted for? (Model development studies only)

As discussed under signalling questions 4.1, 4.2 and 4.5, quantifying the predictive performance of a model on the same data from which the model was developed (apparent performance) tends to give optimistic estimates of performance due to overfitting, i.e. the model is too much adapted to the development data set. This optimism is higher when any of the following are present: total number of outcome events is small; too few outcome events relative to the number of candidate predictors is present (small EPV); dichotomisation of continuous predictors; predictor selection strategies based on

univariable analyses are used; or traditional stepwise predictor selection strategies (e.g. forwards or backwards selection) in multivariable analysis in small data sets (small EPV) are used.(49, 81)

Therefore, studies developing prediction models should always include some form of internal validation, such as bootstrapping and cross-validation. Internal validation is important to quantify overfitting of the developed model and optimism in its predictive performance, except when sample size and notably EPV are extremely large. Internal validation means that only the data of the original sample are used, i.e. validation is based on the same participant data. If there is optimism then an important further step is to adjust or shrink the model predictive performance estimates (such as c-index) as well as the predictor effects in the final model. Unfortunately, this is rarely done. The use of regression coefficients which have not been shrunk or adjusted for optimism will lead to biased (commonly too extreme) predictions when the unshrunk model is used in other individuals. For example, a uniform (linear) shrinkage factor, as can be obtained from a bootstrap procedure, might be applied to all estimated predictor effects. Penalised regression approaches are also becoming popular, such as ridge regression and Lasso regression, which allow each predictor effect to be shrunk differently and even allow exclusion of some predictors entirely.(199) Some authors suggest there is not much difference across different shrinkage methods,(200, 201) but others argue in favour of penalised approaches.(49, 199)

When developing a prediction model, the need to adjust for model overfitting and optimism is thus greater for studies with a small sample size, low EPV and studies using stepwise predictor selection strategies. When internal validation and shrinkage techniques have been used, this signalling question should be classed as Y. Appropriate adjustments for overfitting alleviate the risk of bias concerns due to the issues of low EPV (signalling question 4.1), dichotomisation of continuous predictors (signalling question 4.2), and predictor selection procedures (signalling question 4.5). Studies that develop a prediction model but do not examine or ignore misfitted models should be rated N for this signalling question, certainly in presence of small samples, low EPV, categorisation of continuous predictors and when predictor selection strategies have been used. An exception would be extremely large development studies with high EPV where overfitting is of limited concern.

Some studies may examine or adjust for optimism but use an inappropriate method. Researchers often randomly split a dataset at the participant level in two (one for model development and one for internal validation) which has been shown to be an inadequate way to measure optimism.(154, 202) Secondly, researchers often apply bootstrapping and cross-validation techniques to examine optimism but fail to replicate the exact same model development procedure (e.g. predictor selection procedures, both in univariable analysis and multivariable analysis) and thus may underestimate the actual optimism for their model.(203, 204) Such inappropriate methods would lead to an N for this signalling question.

Example:
Perel et al. examine the impact of overfitting in their model development by using bootstrapping.(89) The authors state: "We drew 200 samples with replacement from the original data, with the same size as the original derivation data. In each bootstrap sample, we repeated the entire modelling process, including variable selection. We averaged the c-statistics of those 200 models in the bootstrap samples. We then estimated the average c-statistic when each of the 200 models was applied in the original sample. The difference between the two average c-statistics indicated the "optimism" of the c statistic in our prognostic model." However, although the optimism in the c-statistic was examined, there was no consideration of the optimism in absolute risk predictions, and thus no shrinkage factor was applied to the predictor coefficients. Nevertheless, the reported optimism in the c-statistic was very small (0.001), i.e. the signalling question should be answered as PY or Y.
In contrast, Rietveld et al. should be answered as PN or N as statistical methods to address overfitting were not used.(90) The authors used a predictor selection procedure based first on univariable p-values and then on multivariable p-values, and additionally considered interactions between included predictors; thus, there is large potential for overfitting. However, no examination of overfitting was made, and no attempt to shrink due to optimism was reported. The authors do report using bootstrapping. However, this appears to be used as a check on the impact of outliers and estimating confidence intervals, rather than to examine overfitting and optimism in discrimination and calibration performance.

4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? (Model development studies only)

Predictors and coefficients of the final developed model, including intercept or baseline components, should be fully reported to allow others to correctly apply it to other individuals. A mismatch between the presented final model and the reported results from the multivariable analysis (e.g. the intercept and predictor coefficients) is frequent. A review of prediction models in cancer in 2010 identified only 13 out of 38 (34%) of final prediction model equations used the same predictors and coefficients of the final presented multivariable analyses, 8 used the same predictors but different coefficients, 11 used neither the same coefficients nor predictors, and in 6 the method to derive the final prediction model from the presented results of the multivariable analysis was unclear.(122)

Bias can arise when there is a mismatch between the presented final model and the results reported from the multivariable analysis. One way in which this can occur is the problem of dropping non-significant predictors from a larger model to arrive at a final presented model but using the predictor coefficients from the larger model which are no longer correct. When dropping predictors from a larger model it is important to re-estimate all predictor coefficients of the smaller model as this has become the final model. These newly estimated predictor coefficients are likely different even if non-significant or non-relevant predictors from the larger model are dropped.

When the study reports the final model where both the predictors and the regression coefficients correspond to the reported results of the multivariable regression analysis or model, then this should be answered as Y. If the final model presented is only based on a selection of predictors from the reported multivariable regression analysis without refitting the smaller model, then this should be answered as N or PN. When there is no information on the multivariable modelling where the predictors and regression coefficients are derived from, then this should be answered as NI.

This signalling question is not about detecting improper methods of selecting predictors for the final model; methods of selecting predictors is addressed in signalling question 4.5.

Example:
Perel et al. report the final model with odds ratios for each predictor and interaction term, and the model formula with predictor coefficients. The full model would be rated as PY or Y as all predictors from the final multivariable analysis are included with coefficients derived from the multivariable analysis. Perel et al. also include a simplified model that was separately developed and validated, with the coefficient terms refitted in the simplified model. If instead the simplified model had not been refitted to correct coefficients for this simplified model with fewer predictors, the paper would have been answered as N for this signalling question. In Rietveld et al., all predictors in the final model were included in the simplified clinical score but this simplified clinical score used whole number scores, presumably to facilitate its usability. However, these rounded number scores no longer weighted the predictors based on the final model, as seen for the predictor "two glued eyes" where the coefficient of 2.707 was rounded to 5 (multiplied by 1.84), whereas -0.61 was rounded to -1 (multiplied by 1.64). The signalling question would be answered N as the assigned weights of the predictors do not correspond to the results in the final multivariable analysis.

Rating the risk of bias for domain 4

Table 10 shows how the signalling questions should be answered and an overall judgement for domain 4 should be reached.

Tailoring PROBAST with additional signalling questions

We encourage researchers to also use PROBAST to appraise prediction model studies in which other outcome types than binary or time-to-event outcomes (e.g. for ordinal, nominal or continuous outcomes) were considered, and for studies using alternative analysis methods to regression-based techniques (e.g. tree based, machine or artificial learning techniques). Reviewers may tailor PROBAST by adding additional signalling questions to address bias related to these other types of outcomes or modelling techniques. For example, when addressing models for prediction of continuous outcomes, the signalling question that addresses the number of events per studied predictor (Domain 4) may be tailored to address the total number of study participants per studied predictor.(49) When studies based on machine or artificial learning techniques are used, most if not all of the signalling questions will still apply. Additional questions may need to be added, as these techniques use different predictor selection strategies, predictor-outcome estimations and methods to adjust for overfitting.

Also, when investigating studies on the added predictive value of a specific predictor to an existing model, a signalling question can be added that focuses on the methods used for quantifying added value, for example net reclassification index (NRI) or decision curve analysis.(85, 205) Similarly, when investigating studies that focus on recalibration or updating an existing model to another setting, a question on the method of recalibration or updating could be added, for example recalibrating the baseline risk or hazard, updating the original regression coefficients, or refitting the entire model.

Whenever reviewers decide to tailor or add signalling questions, these need to be phrased such that the answer "yes" indicates a low risk of bias, to facilitate coherence with current signalling questions. Specific guidance on how to assess each added signalling question specific for a review should also be produced.

We do not recommend removing signalling questions from the tool unless they are clearly not relevant to a review question. If all studies would rate "yes" or "no" for a particular question, then it is still helpful to leave it in the tool. This shows whether a particular source of bias or concern for applicability is a potential problem for that review.

Step 4 – Overall judgement

Table 11 shows an overall judgement on the risk of bias and applicability of a prediction model evaluation. If a prediction model evaluation is judged as "low" on all domains relating to bias or applicability, then it is appropriate to have an overall judgment of "low risk of bias" or "low concern for applicability". If an evaluation is judged "high" for at least one domain, then it should be judged at "high risk of bias" or as having "high concerns for applicability." If the prediction model evaluation is "unclear" in one or more domains and was rated as "low" in the remaining domains, then it may be judged at "unclear risk of bias" or as having "unclear concerns for applicability".

PROBAST should not be used to generate a summary "quality score" for a study because of the well-known problems associated with such scores.(206, 207) Rather than striving for a summary score, the impact of problems within each domain should be judged and discussed.

Presentation and use of PROBAST assessment into the review

Presentation of the risk of bias and applicability assessment is an important aspect of communicating the strength of evidence in a review. All reviews should include a narrative summary of risk of bias and applicability concerns, linked to how this affects interpretation of findings and strength of inferences. In addition, a table showing the results of the assessments of risk of bias and applicability concerns of all included assessments should be presented. Table 12 is an example to facilitate identification of key issues across all included prediction models and their studies. A quick way to summarise across all studies is a graphical summary presenting the percentage of studies rated by level of concern risk of bias and applicability for each domain (see Figure 1). This is in line with item 22 of the PRISMA statement of how to report systematic reviews and meta-analyses of studies that evaluate health care interventions (PRISMA).(39, 40) It should be noted that these summaries are not sufficient on their own, i.e. without an accompanying discussion of what any observed patterns mean for the evidence base in relation to the review question.

Further incorporation of risk of bias and concerns for applicability may be specified in the review planning stage or in the systematic review protocol. Findings can be included in the analysis by planning sensitivity analyses limited to studies with low concerns for risk of bias or applicability either overall or for particular domains, or investigation of heterogeneity between studies using subgroups based on ratings of concern.(20)

Concluding remarks

(308 words)

PROBAST is the first rigorously developed tool designed specifically to assess the risk of bias and concerns for applicability of primary studies on development, validation or updating (including extension) of prediction models to be used for individualised predictions. PROBAST covers both diagnostic and prognostic models, regardless of the medical domain, type of outcome, predictors or statistical technique used.

This E&E paper provides explicit guidance on how to use PROBAST (REF M18-1376), including how to interpret each signalling question, how to grade the risk of bias per domain and overall, and how to present and incorporate PROBAST assessments in a systematic review, all accompanied with generic guidance on diagnostic and prognostic prediction model research. This detailed explanation and elaboration for PROBAST will enable a focussed and transparent approach to assessing the risk of bias and applicability of studies developing, validating or updating of prediction models for individualised predictions. Six worked-out examples of PROBAST assessments, covering development studies, validation studies, a combination of both and addressing both diagnostic and prognostic models can be found at our website www.probast.org. We also encourage and will make available translations of PROBAST.

The use of PROBAST requires expertise and knowledge of prediction model researchers as well as clinicians. Guidance on methods for prediction model research is still at an early stage compared to guidance on methods and interpretation of randomised intervention studies and diagnostic test accuracy studies. We recognise that currently necessary information for assessment of bias and applicability is often not reported, and hope that adherence of both journals and authors to the TRIPOD reporting guideline (7, 8) will reduce this problem.

As with other risk of bias and reporting guidelines in medical research, PROBAST and its guidance will require updating, as methods for prediction model studies develop. We recommend downloading the latest version of PROBAST tool and guidance from the website (www.probast.org).

Contact details for all authors

Karel G. M. Moons
Julius Centre for Health Sciences and Primary Care
UMC Utrecht
Utrecht University
PO Box 85500
3508 GA Utrecht
The Netherlands
K.G.M.Moons@umcutrecht.nl

Robert F. Wolff
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
robert@systematic-reviews.com

Richard D. Riley
Centre for Prognosis Research,
Research Institute for Primary Care and Health Sciences
Keele University
Staffordshire ST5 5BG
United Kingdom
r.riley@keele.ac.uk

Penny F. Whiting
NIHR CLAHRC West
University Hospitals Bristol NHS Foundation Trust, Bristol, United Kingdom
School of Social and Community Medicine, University of Bristol, United Kingdom
Whitefriars BS1 2NT
United Kingdom
Penny.Whiting@bristol.ac.uk

Marie Westwood
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
marie@systematic-reviews.com

Gary S. Collins
Centre for Statistics in Medicine, NDORMS, University of Oxford
Botnar Research Centre, Windmill Road

Oxford OX3 7LD
United Kingdom
gary.collins@csm.ox.ac.uk

Johannes B. Reitsma
Julius Centre for Health Sciences and Primary Care
UMC Utrecht
Utrecht University
PO Box 85500
3508 GA Utrecht
The Netherlands
J.B.Reitsma-2@umcutrecht.nl

Jos Kleijnen
Kleijnen Systematic Reviews Ltd, Unit 6, Escrick Business Park, Riccall Road, Escrick, York YO19 6FD, United Kingdom
School for Public Health and Primary Care (CAPHRI) Maastricht University, Maastricht, The Netherlands
jos@systematic-reviews.com

Sue Mallett
Institute of Applied Health Sciences
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom
s.mallett@bham.ac.uk

Acknowledgement

The authors would like to thank the members of the Delphi panel (see below) for their valuable input. Furthermore, the authors would like to thank all testers, especially Cordula Braun, Johanna A.A.G. Damen, Paul Glasziou, Pauline Heus, Lotty Hooft, and Romin Pajouheshnia, for providing feedback on PROBAST. The authors are grateful to Janine Ross and Steven Duffy for their support in managing the references.

KGM Moons and JB Reitsma gratefully acknowledges financial contribution by the Netherlands Organisation for Scientific Research (ZONMW 918.10.615 and 91208004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

R Riley is a member of the Evidence Synthesis Working Group funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [ProjectNumber 390]. The views expressed are those of the author(s) and not necessarily those of the NIHR, the NHS or the Department of Health.

PF Whiting (time) was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) West at University Hospitals Bristol NHS Foundation Trust.

GS Collins was supported by the NIHR Biomedical Research Centre, Oxford.

S Mallett is supported by NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

This report presents independent research supported by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, or the Department of Health.

Potential conflicts of interest

Karel G. M. Moons: None to declare

Robert F. Wolff: None to declare

Richard D. Riley: None to declare

Penny F. Whiting: None to declare

Marie Westwood: None to declare

Gary S. Collins: None to declare

Johannes B. Reitsma: None to declare

Jos Kleijnen: None to declare

Sue Mallett: None to declare

Author Contributions

Conception and design: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Analysis and interpretation of the data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Drafting of the article: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Critical revision for important intellectual content: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Final approval of the article: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Statistical expertise: K.G.M. Moons, R.D. Riley, G.S. Collins, J.B. Reitsma, S. Mallett

Obtaining of funding: K.G.M. Moons, R.D. Riley, P.F. Whiting, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Administrative, technical, or logistic support: R.F. Wolff, K.G.M. Moons, J. Kleijnen

Collection and assembly of data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

References

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
2. Harrell FE, Jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. Journal of the National Cancer Institute. 1988;80(15):1198-202.
3. Hlatky MA. Evaluation of diagnostic tests. Journal of Chronic Diseases. 1986;39(5):357-60.
4. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277(6):488-94.
5. Sox H, Jr. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. Annals of Internal Medicine. 1986;104(1):60-6.
6. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. New England Journal of Medicine. 1985;313(13):793-9.
7. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
8. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-73.
9. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, eds. Breast cancer Translational therapeutic strategies. New York: Informa Healthcare; 2007:11-25.
10. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. American Journal of Obstetrics and Gynecology. 2016;214(1):79-90 e36.
11. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical prediction models for cardiovascular disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circulation. Cardiovascular Quality and Outcomes. 2015;8(4):368-75.
12. Murad MH, Montori VM. Synthesizing evidence: shifting the focus from individual studies to the body of evidence. JAMA. 2013;309(21):2217-8.
13. Hemingway H. Prognosis research: why is Dr. Lydgate still waiting? Journal of Clinical Epidemiology. 2006;59(12):1229-38.
14. Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. Journal of Clinical Epidemiology. 2007;60(8):863-5; author reply 5-6.
15. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. Journal of the American Medical Informatics Association. 2001;8(4):391-7.
16. Keogh C, Wallace E, O'Brien KK, Murphy PJ, Teljeur C, McGrath B, et al. Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. Journal of Clinical Epidemiology. 2011;64(8):848-60.
17. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA ... Annual Symposium proceedings. AMIA Symposium. 2003:728-32.
18. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7(2):e32844.
19. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Medicine. 2014;11(10):e1001744.

20. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
21. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
22. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. Journal of Clinical Epidemiology. 2016;69:40-50.
23. Deeks JJ, Wisniewski S, Davenport C**.** Chapter 4: guide to the contents of a Cochrane diagnostic test accuracy protocol. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy: The Cochrane Collaboration; 2013.
24. Bossuyt PM, Leeflang MM**.** Chapter 6: developing criteria for including studies. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy: The Cochrane Collaboration; 2008.
25. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D**.** Chapter 7: searching for studies. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy: The Cochrane Collaboration; 2008.
26. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ**.** Chapter 9: assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy: The Cochrane Collaboration; 2009.
27. Ahmed I, Debray TP, Moons KG, Riley RD**.** Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Medical Research Methodology. 2014;14:3.
28. Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KG**.** Meta-analysis and aggregation of multiple published prediction models. Statistics in Medicine. 2014;33(14):2341-62.
29. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW**.** Aggregating published prediction models with individual participant data: a comparison of different approaches. Statistics in Medicine. 2012;31(23):2697-712.
30. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD**.** Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? Statistical Methods in Medical Research. 2017:962280217705678.
31. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y**.** Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy: The Cochrane Collaboration; 2010.
32. Chu H, Guo H, Zhou Y**.** Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. Medical Decision Making. 2010;30(4):499-508.
33. Dendukuri N, Hadgu A, Wang L**.** Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Statistics in Medicine. 2009;28(3):441-61.
34. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA**.** A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.
35. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH**.** Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005;58(10):982-90.
36. Rutter CM, Gatsonis CA**.** A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine. 2001;20(19):2865-84.
37. Takwoingi Y, Guo B, Riley RD, Deeks JJ**.** Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Statistical Methods in Medical Research. 2017;26(4):1896-911.
38. Takwoingi Y, Leeflang MM, Deeks JJ**.** Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Annals of Internal Medicine. 2013;158(7):544-54.

39. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
40. Moher D, Liberati A, Tetzlaff J, Altman DG**.** Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.
41. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA Statement. JAMA. 2018;319(4):388-96.
42. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology. 2016;69:225-34.
43. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell, The Cochrane Collaboration; 2011.
44. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
45. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, eds. Cochrane Methods; 2016.
46. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;155(8):529-36.
47. Canet J, Gallart L, Gomar C, Paluzie G, Valles J, Castillo J, et al. Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010;113(6):1338-50.
48. Collins GS, Omar O, Shanyinde M, Yu LM**.** A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of Clinical Epidemiology. 2013;66(3):268-77.
49. Harrell FE**.** Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
50. Royston P, Moons KG, Altman DG, Vergouwe Y**.** Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
51. Lamain-de Ruiter M, Kwee A, Naaktgeboren CA, de Groot I, Evers IM, Groenendaal F, et al. External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. BMJ. 2016;354:i4338.
52. Hippisley-Cox J, Coupland C**.** Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. BMJ. 2012;344:e3427.
53. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793.
54. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine. 2013;10(2):e1001381.
55. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
56. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.
57. van der Ploeg T, Austin PC, Steyerberg EW**.** Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014;14:137.
58. Collins GS, Mallett S, Omar O, Yu LM**.** Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9:103.

59. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. Cerebrovascular Diseases. 2001;12(3):159-70.
60. Tamariz LJ, Eng J, Segal JB, Krishnan JA, Bolger DT, Streiff MB, et al. Usefulness of clinical prediction rules for the diagnosis of venous thromboembolism: a systematic review. American Journal of Medicine. 2004;117(9):676-84.
61. Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW. Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke. 2011;42(5):1482-8.
62. Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, et al. Prediction models in reproductive medicine: a critical appraisal. Human Reproduction Update. 2009;15(5):537-52.
63. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Medical Informatics and Decision Making. 2006;6:38.
64. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. European Journal of Cardio-thoracic Surgery. 2012;41(4):746-54.
65. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318.
66. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302(21):2345-52.
67. Peters SA, den Ruijter HM, Bots ML, Moons KG. Improvements in risk stratification for the occurrence of cardiovascular disease by imaging subclinical atherosclerosis: a systematic review. Heart. 2012;98(3):177-84.
68. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Medicine. 2012;9(5):1-12.
69. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Medicine. 2013;10(2):e1001380.
70. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013;158(4):280-6.
71. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
72. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). BMC Medical Informatics and Decision Making. 2011;11:62.
73. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of Internal Medicine. 2006;144(3):201-9.
74. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of Internal Medicine. 1999;130(6):515-24.
75. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
76. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in Medicine. 2014;33(3):517-35.
77. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167-76.
78. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine. 2016;35(2):214-26.
79. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Statistical Methods in Medical Research. 2016;25(4):1692-706.

80. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
81. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
82. Grønnesby JK, Borgan O. A method for checking regression models in survival analysis based on the risk score. Lifetime Data Analysis. 1996;2(4):315-28.
83. D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. Handbook of statistics, survival methods. Amsterdam: Elsevier; 2004.
84. Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry. 2008;54(4):729-37.
85. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-74.
86. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Annals of Internal Medicine. 2005;143(2):100-7.
87. Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A. Development of a cardiovascular risk score for use in low- and middle-income countries. Journal of Nutrition. 2011;141(7):1375-80.
88. Han J, King NK, Neilson SJ, Gandhi MP, Ng I. External validation of the CRASH and IMPACT prognostic models in severe traumatic brain injury. Journal of Neurotrauma. 2014;31(13):1146-52.
89. Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012;345:e5166.
90. Rietveld RP, ter Riet G, Bindels PJ, Sloos JH, van Weert HC. Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. BMJ. 2004;329(7459):206-10.
91. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology. 2015;44(3):827-36.
92. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. Journal of Clinical Epidemiology. 2016;78:90-100.
93. Schuit E, Groenwold RH, Harrell FE, Jr., de Kort WL, Kwee A, Mol BW, et al. Unexpected predictor-outcome associations in clinical prediction research: causes and solutions. CMAJ. 2013;185(10):E499-505.
94. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279-89.
95. van Klaveren D, Gonen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Statistics in Medicine. 2016;35(23):4136-52.
96. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG. Adaptation of clinical prediction models for application in local settings. Medical Decision Making. 2012;32(3):E1-10.
97. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American Journal of Epidemiology. 2010;172(8):971-80.
98. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. American Journal of Epidemiology. 2012;175(7):715-24.

99. Kengne AP, Beulens JWJ, Peelen LM, Moons KGM, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. The Lancet Diabetes & Endocrinology. 2014;2(1):19-29.
100. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice - experiences from the MORGAM Project. Epidemiologic Perspectives and Innovations : EP+I. 2007;4:15.
101. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Medical Research Methodology. 2013;13:113.
102. Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. London: Jones and Bartlett Publishers; 2009.
103. Knottnerus JA. The evidence base of clinical diagnosis. London: BMJ Books; 2002.
104. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of Clinical Epidemiology. 2003;56(11):1118-28.
105. Sackett DL, Tugwell P, Guyatt GH. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston: Little, Brown & Co; 1991.
106. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Medical Research Methodology. 2008;8:48.
107. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061-6.
108. Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JP, Hernández-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clinical Chemistry. 2009;55(4):786-94.
109. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clinical Chemistry. 2005;51(8):1335-41.
110. van Zaane B, Vergouwe Y, Donders AR, Moons KG. Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. BMC Medical Research Methodology. 2012;12:166.
111. van der Leeuw J, van Dieren S, Beulens JW, Boeing H, Spijkerman AM, van der Graaf Y, et al. The validation of cardiovascular risk scores for patients with type 2 diabetes mellitus. Heart. 2015;101(3):222-9.
112. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. Radiology. 1988;167(2):565-9.
113. Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine. 1987;6(4):411-23.
114. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. JAMA. 1997;277(1):49-52.
115. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clinical Radiology. 1995;50(8):513-8.
116. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? Journal of Clinical Epidemiology. 2002;55(7):633-6.
117. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. Journal of Clinical Epidemiology. 2013;66(10):1093-104.
118. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994;271(9):389-91.
119. Schwartz W, Wolfe HJ, Pauker SG. Pathology and probabilities: a new approach to interpreting and reporting biopsies. New England Journal of Medicine. 1981;305(16):917-23.
120. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240(4857):1285-93.
121. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174(4):469-76.

122. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Medicine. 2010;8:20.
123. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. British Journal of Cancer. 1994;69(6):979-85.
124. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. Journal of Clinical Epidemiology. 2009;62(8):797-806.
125. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard--a systematic review. American Journal of Epidemiology. 2014;179(4):423-31.
126. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statistics in Medicine. 1999;18(22):2987-3003.
127. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Statistical Methods in Medical Research. 1998;7(4):354-70.
128. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology. 1999;10(1):67-72.
129. Lu J, Marmarou A, Lapane KL. Impact of GOS misclassification on ordinal outcome analysis of traumatic brain injury clinical trials. Journal of Neurotrauma. 2012;29(5):719-26.
130. Lu J, Murray GD, Steyerberg EW, Butcher I, McHugh GS, Lingsma H, et al. Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. Journal of Neurotrauma. 2008;25(6):641-51.
131. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. Epidemiology. 1992;3(5):434-40.
132. Naaktgeboren CA, Bertens LC, van Smeden M, Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605.
133. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown & Co; 1985.
134. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders Company; 1985.
135. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine. 1978;299(17):926-30.
136. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Medicine. 2013;10:e1001531.
137. Naaktgeboren CA, de Groot JA, van Smeden M, Moons KG, Reitsma JB. Evaluating diagnostic accuracy in the face of multiple reference standards. Annals of Internal Medicine. 2013;159(3):195-202.
138. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770.
139. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, et al. Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. American Journal of Epidemiology. 2012;175(8):847-53.
140. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39(1):207-15.
141. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Statistics in Medicine. 2006;25(22):3769-86.
142. de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. Statistics in Medicine. 2008;27(28):5880-9.
143. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment. 2007;11(50):iii, ix-51.

144. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. (in press). Minimum sample size for developing a multivariable prediction model: PART I - continuous outcomes. Stat Med. 2018.
145. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. (in press). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2018.
146. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology. 2016;16(1):163.
147. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG**.** Internal and external validation of predictive models: a simulation study of bias and precision in small samples. Journal of Clinical Epidemiology. 2003;56(5):441-7.
148. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD**.** Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical Decision Making. 2001;21(1):45-56.
149. Peduzzi P, Concato J, Feinstein AR, Holford TR**.** Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology. 1995;48(12):1503-10.
150. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR**.** A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology. 1996;49(12):1373-9.
151. Vittinghoff E, McCulloch CE**.** Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology. 2007;165(6):710-8.
152. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV**.** Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. Journal of Clinical Epidemiology. 2011;64(9):993-1000.
153. Ogundimu EO, Altman DG, Collins GS**.** Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. 2016;76:175-82.
154. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD**.** Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology. 2001;54(8):774-81.
155. Altman DG, Vergouwe Y, Royston P, Moons KG**.** Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
156. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
157. Altman DG, Royston P**.** What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19(4):453-73.
158. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD**.** Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of Clinical Epidemiology. 2005;58(5):475-83.
159. Altman DG, Lausen B, Sauerbrei W, Schumacher M**.** Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute. 1994;86(11):829-35.
160. Altman DG, Royston P**.** The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.
161. Royston P, Altman DG, Sauerbrei W**.** Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine. 2006;25(1):127-41.
162. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG**.** Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in Medicine. 2016;35(23):4124-35.

163. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods. 2002;7(1):19-40.
164. Royston P, Sauerbrei W. Multivariable model-building - a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley; 2008.
165. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. Journal of Chronic Diseases. 1986;39(8):575-84.
166. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013;346:f2778.
167. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: Wiley; 2002.
168. Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.
169. Schafer JL. Multiple imputation: a primer. Statistical Methods in Medical Research. 1999;8(1):3-15.
170. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine. 1999;18(6):681-94.
171. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99.
172. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006;59(10):1087-91.
173. Janssen KJ, Donders AR, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. Journal of Clinical Epidemiology. 2010;63(7):721-7.
174. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Medical Research Methodology. 2010;10:7.
175. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
176. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology. 2010;63(2):205-14.
177. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184(11):1265-9.
178. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory heart failure patients: a systematic review. Circulation. Heart Failure. 2013;6(5):881-9.
179. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Investigation. 2009;27(3):235-43.
180. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14:40.
181. Hussain A, Dunn KW. Predicting length of stay in thermal burns: A systematic review of prognostic factors. Burns. 2013;39:1331-40.
182. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Research and Treatment. 2012;132(2):365-77.
183. Medlock S, Ravelli ACJ, Tamminga P, Mol BWM, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6:e23441.
184. Steurer J, Haller C, Häuselmann H, Brunner F, Bachmann LM. Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. PLoS One. 2011;6(5):e19994.

185. van Dijk WD, Bemt L, Haak-Rongen S, Bischoff E, Weel C, Veen JC, et al. Multidimensional prognostic indices for use in COPD patient care. A systematic review. Respiratory Research. 2011;12:151.
186. Vuong K, McGeechan K, Armstrong BK, Cust AE. Risk prediction models for incident primary cutaneous melanoma: a systematic review. JAMA Dermatology. 2014;150(4):434-44.
187. Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006;59(10):1092-101.
188. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology. 2009;9:57.
189. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clinical Chemistry. 2009;55(5):994-1001.
190. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Statistics in Medicine. 2015;34(11):1841-63.
191. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology. 1996;49(8):907-16.
192. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicne. 1996;15(4):361-87.
193. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555-61.
194. Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. Statistics in Medicine. 2014;33(22):3844-58.
195. Gail MH, Wieland S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984;71:431e44.
196. Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. Statistical Science. 1999;14(1):29-46.
197. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Statistical Methods in Medical Research. 2016;27(6):1723-36.
198. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. Journal of Clinical Epidemiology. 2006;59(8):798-801.
199. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868.
200. Janssen KJ, Siccama I, Vergouwe Y, Koffijberg H, Debray TP, Keijzer M, et al. Development and validation of clinical prediction models: marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming. Journal of Clinical Epidemiology. 2012;65(4):404-12.
201. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000;19(8):1059-79.
202. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical Methods in Medical Research. 2014;26(6):796-808.
203. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Briefings in Bioinformatics. 2011;12(3):189-202.
204. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.

205. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine. 2011;30(1):11-21.
206. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282(11):1054-60.
207. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Medical Research Methodology. 2005;5:19.

Appendix

Members of PROBAST steering group

Karel G. M. Moons, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Robert F. Wolff, Kleijnen Systematic Reviews, York, United Kingdom
Richard D. Riley, Keele University, United Kingdom
Penny F. Whiting, University Hospitals Bristol NHS Foundation Trust, United Kingdom; University of Bristol, United Kingdom
Marie Westwood, Kleijnen Systematic Reviews, York, United Kingdom
Gary S. Collins, Centre for Statistics in Medicine, NDORMS, University of Oxford, United Kingdom
Johannes B. Reitsma, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Jos Kleijnen, Kleijnen Systematic Reviews, York, United Kingdom; School for Public Health and Primary Care (CAPHRI), Maastricht University, Maastricht, The Netherlands
Sue Mallett, Institute of Applied Health Sciences, University of Birmingham, United Kingdom

Members of PROBAST Delphi group (in alphabetical order)

Prof Doug Altman, PhD. Centre for Statistics in Medicine, NDORMS, University of Oxford, United Kingdom
Prof Patrick Bossuyt, PhD. Division Clinical Methods & Public Health, University of Amsterdam, The Netherlands
Prof Nancy R. Cook, ScD. Brigham and Women's Hospital, Boston, United States of America
Gennaro D´Amico, MD. Ospedale V Cervello, Palermo, Italy
Thomas P. A. Debray, PhD, MSc. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Jon Deeks, PhD. Institute of Applied Health Research, University of Birmingham, United Kingdom
Joris de Groot, PhD. Philips Image Guided Therapy Systems, Best, The Netherlands
Emanuele di Angelantonio, PhD, MSc. Department of Public Health and Primary Care, University of Cambridge, United Kingdom
Prof Tom Fahey, MD, MSc. Royal College of Surgeons in Ireland, Dublin, Ireland
Prof Frank Harrell, PhD. Department of Biostatistics, Vanderbilt University, United States of America
Prof Jill A. Hayden, PhD. Department of Community Health and Epidemiology, Dalhousie University, Canada
Martijn W. Heymans, PhD. Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands
Lotty Hooft, PhD. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Chris Hyde, PhD. Institute of Health Research, University of Exeter Medical School, United Kingdom
Prof John Ioannidis, MD, DSc. Meta-Research Innovation Center at Stanford (METRICS), Stanford University, United States of America
Prof Alfonso Iorio, MD, PhD. Department of Health Research Methods, Evidence, and Impact (HEI), McMaster University, Canada
Stephen Kaptoge, PhD. Department of Public Health & Primary Care, University of Cambridge, United Kingdom
Prof André Knottnerus, MD, PhD. Department of Family Medicine, Maastricht University, The Netherlands
Mariska Leeflang, PHD, DVM. Department of Clinical Epidemiology, Biostatistics and Boinformatics, University of Amsterdam, The Netherlands
Frances Nixon, BSc. National Institute for Health and Care Excellence (NICE), Manchester, United Kingdom
Prof Pablo Perel, MD, PhD, MSc. Centre for Global Chronic Conditions, London School of Hygiene and Tropical Medicine, United Kingdom
Bob Phillips, PhD, MMedSci. Centre for Reviews and Dissemination (CRD), York, United Kingdom
Heike Raatz, MD, MSc. Kleijnen Systematic Reviews, York, United Kingdom
Rob Riemsma, PhD. Kleijnen Systematic Reviews, York, United Kingdom
Prof Maroeska Rovers, PhD. Departments of Operating Rooms and Health Evidence, Radboud University Medical Center, Nijmegen, The Netherlands
Anne W. S. Rutjes, PhD, MHSc. Institute for Social and Preventive Medicine (ISPM) and Institute of Primary Health Care (BIHAM), University of Bern, Switzerland
Prof Willi Sauerbrei, PhD. Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany
Stefan Sauerland, MD, MPH. Institute for Quality and Efficiency in Healthcare (IQWiG), Cologne, Germany
Fülöp Scheibler, PhD, MA. University Medical Center Schleswig-Holstein, Kiel, Germany
Prof Rob Scholten, MD, PhD. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Ewoud Schuit, PhD, MSc. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Ewout Steyerberg, PhD. Department of Public Health, Erasmus University Medical Center Rotterdam and Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands
Toni Tan, MSc. National Institute for Health and Care Excellence (NICE), Manchester, United Kingdom
Gerben ter Riet, MD, PhD. Department of General Practice, University of Amsterdam, The Netherlands
Prof Danielle van der Windt, PhD. Centre for Prognosis Research, Keele University, United Kingdom
Yvonne Vergouwe, PhD. Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands
Andrew Vickers, PhD. Memorial Sloan-Kettering Cancer Center, New York, United States of America
Angela M. Wood, PhD. Department of Public Health and Primary Care, University of Cambridge, United Kingdom

Tables

Table 1. Guidance on conducting systematic reviews of prediction model studies

Table 2. PICOTS

Six key items (the so-called PICOTS) to guide the framing of the review aim. PICOTS is a modification of the traditional PICO system used in systematic reviews of therapeutic intervention studies, by adding Timing (the time point of using the prediction model and the time period of the prediction) and clinical Setting(19, 20)

Table 3. Four steps in PROBAST

Table 4. Example papers

Table 5. Example Step 1 applied to the Perel example study(89)

Table 6. Example Step 2 applied to the Perel example study(89)

Table 7. Participants domain: guidance notes for rating risk of bias and applicability

Table 8. Predictors domain: guidance notes for rating risk of bias and applicability

Table 9. Outcome domain: guidance notes for rating risk of bias and applicability

Table 10. Analysis domain: guidance notes for rating risk of bias

Table 11. Overall assessment of risk of bias and concerns for applicability

Table 12. Suggested Tabular Presentation for PROBAST Results

Figure

Figure 1. Suggested Graphical Presentation for PROBAST Results

Boxes

Box 1. Types of diagnostic and prognostic modelling studies or reports addressed by PROBAST

(adopted from the TRIPOD and CHARMS guidance(8, 19))

Box 2. Differences between diagnostic and prognostic prediction model studies

Box 3. Examples of systematic review questions for which PROBAST is suitable

Box 4. Prediction model performance measures

PROBAST: a tool to assess risk of bias and applicability of prediction model studies – explanation and elaboration

Table 1. Guidance on conducting systematic reviews of prediction model studies

Task	Guidance
Reporting of primary study	Transparent reporting of prediction models for prognosis and diagnosis (TRIPOD)(7, 8)
Defining review question and developing criteria for including studies*	Guidance for defining review question and design of the review of prognosis studies , see Table 4 (CHARMS)(19) (20) Guidance for protocol for diagnostic test accuracy (DTA) reviews(23, 24)
Searching for studies*	Search filters for prediction studies(18) https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/filters-to-identify-studies-about-prognosis Search for DTA studies(25)
Selecting studies and extracting data*	Guidance and checklist for data extraction and critical appraisal of prognosis studies (CHARMS)(19). Guidance for DTA studies(24, 26)
Assessing risk of bias and applicability in included studies*	Prediction model Risk Of Bias Assessment Tool (PROBAST)[REF M18-1376]
Analysing data and undertaking meta-analyses*	Meta-analysis of prediction models(20, 27-30); Meta-analysis of diagnostic test accuracy studies(31-38)
Interpreting results and drawing conclusions*	PROBAST [REF M18-1376] Guidance for interpretation of results(20, 27-29) Guidance for interpretation of diagnostic test accuracy studies(24)
Reporting of systematic reviews	Transparent reporting of systematic reviews and meta-analysis (PRISMA)(39-41)
Assessing risk of bias of systematic reviews	Risk of bias in systematic reviews (ROBIS)(42)

* Step in line with the general methods for Cochrane Reviews(43)

Table 2. PICOTS

Six key items (the so-called PICOTS) to guide the framing of the review aim. PICOTS is a modification of the traditional PICO system used in systematic reviews of therapeutic intervention studies, by adding Timing (the time point of using the prediction model and the time period of the prediction) and clinical Setting(19, 20)

Item	Comments
1. Population	Define the target population in which the prediction model(s) under review will be used.
2. Index	Define the prediction model(s) under review.
3. Comparator	If applicable, define whether other prediction models are reviewed and compared to the index model
4. Outcome(s)	Define the outcome(s) of interest for the model(s) under review.
5. Timing	Define at what moment or time-point (e.g. in the patient work-up) the prediction model(s) under review are to be used in the targeted population, and over what time period the outcome(s) are predicted (the latter in case of prognostic models).
6. Setting	Define the intended clinical setting of the prediction model(s) under review.

Table 3. Four steps in PROBAST

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability (per domain)	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment of risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication

Table 4. Example papers

Author (Year)	Topic area	Type of model Dev/Val	prediction Diag/Prog	Data source	Study population	Type of predictors	Outcome	Sample size (N outcome events)	Performance Discr.	Cal.
Aslibekyan 2011(87)	Myocardial infarction	Dev+Val	Prog	Non-nested case-control study, population of central valley in Costa Rica (1994-2004)	First non-fatal acute MI	History taking, physical examination	Fist non-fatal MI	4547 (1984)	Yes	No
Han 2014(88)	Severe traumatic brain injury	Val	Prog	Cohort study, 1 hospital in Singapore (02/2006-12/2009)	Severe TBI (GCS≤8)	History taking, physical examination, laboratory parameters, CT	Mortality (14 day, 6 months), unfavourable events (6 months)	300 (143/ 162/ 213)	Yes	Yes
Oudega 2005(86)	Deep vein thrombosis	Val	Diag	Prospective cross-sectional study, 110 primary care practices in the Netherlands (Val: 01/2002 – 03/2003)	Symptomatic DVT	History taking, physical examination	DVT	Val: 1295 (289)	No	No
Perel 2012(89)	Traumatic bleeding	Dev+Val	Prog	Dev: Randomised controlled trial, 274 hospitals in 40 countries (no dates reported)	Trauma or risk of significant bleeding	History taking, type of injury, physiological examination	Mortality	Dev: 20127	Yes	Yes
				Val: Registry, 60% of trauma hospitals in England and Wales (2000-2008)	Blood loss ≥20%			Val: 14220	Yes	Yes
Rietveld 2004(90)	Infectious conjunctivitis	Dev	Diag	Cohort study, 25 care centres in NL (09/1999-12/2002)	Red eye + (muco-) purulent discharge or glued eyelid	History taking, physical examination	Positive bacterial culture	184 (57)	Yes	Yes

Cal = Calibration; Dev = Development; Diag = Diagnostic; Discr. = Discrimination; DVT = deep vein thrombosis; GCS = Glasgow Coma Scale; MI = Myocardial infarction; NL = The Netherlands; Prog = Prognostic; Ref = Refinement; TBI = Traumatic brain injury; Val = Validation

Table 5. Example Step 1 applied to the Perel example study(89)

Criteria	Specify your systematic review question:
Intended use of model:	Prognosis; At presentation at hospital accident and emergency
Participants *including selection criteria and setting:*	Trauma patients presenting at accident and emergency.
Predictors (used in modelling) *including (1) types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), (2) time of measurement, (3) specific measurement issues (e.g. any requirements/ prohibitions for specialised equipment):*	Patients' demographics; Physiological variables; Injury characteristics; Time from injury -- all measured at presentation to A&E. Imaging with results available within 4 hours of admission Key predictors to include: type of injury
Outcome *to be predicted:*	Death within 4 weeks of injury

Table 6. Example Step 2 applied to the Perel example study(89)

Type of prediction study	PROBAST boxes to complete	Tick as appropriate	Definitions for type of prediction model study
Development only	Dev		Prediction model development without external validation. These studies may include internal validation methods such as bootstrapping and cross-validation techniques
Development and validation	Dev and Val	✓	Prediction model development combined with external validation in other participants in the same article
Validation only	Val		External validation of existing (previously developed) model in other participants

Table 7. Participants domain: guidance notes for rating risk of bias and applicability

Domain 1: Participants	
Risk of bias assessment	
Background: The overall aim for prediction models is to generate absolute risk predictions that are correct in new individuals. Certain data sources or designs are not suited to generate absolute probabilities. Problems may also arise if a study inappropriately includes or excludes participant groups from entering the study.	
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	
Yes/ Probably yes or	If a cohort design (including RCT or proper registry data) or a nested case-control case-cohort design (with proper adjustment of the baseline risk/hazard in the analysis) has been used.
No/ Probably no	If a non-nested case-control design has been used.
No information	If the method of participant sampling is unclear.
1.2 Were all inclusions and exclusions of participants appropriate?	
Yes/ Probably yes	If inclusion and exclusion of participants was appropriate, so participants correspond to unselected participants of interest.
No/ Probably no	If participants are included who would already have been identified as having the outcome by prior tests and so are no longer participants at suspicion of disease (diagnostic studies) or at risk of developing outcome (prognostic studies)
the	*or* if specific subgroups are excluded that may have altered the performance of model for the intended target population.
No information place.	When there is no information on whether inappropriate in- or exclusions took
Risk of bias introduced by participants or data sources:	
Low risk of bias	If the answer to all signalling questions is “Yes” or “Probably Yes” then risk of bias can be considered low.
	If one or more of the answers is “No” or “Probably no”, the judgement could still be “Low risk of bias” but specific reasons should be provided why the risk of bias can be considered low.
High risk of bias	If the answer to any of the signalling questions is “No” or “Probably no” there is a potential for bias, except if defined at low risk of bias above.
Unclear risk of bias	If relevant information is missing for some of the signalling questions and none of the signalling questions is judged to put this domain at high risk of bias.
Concerns for applicability	
Background: Included participants, the selection criteria used as well as the setting used in the primary study should be relevant to the review question.	
Concern that included participants or the setting do not match the review question:	
Low concern for applicability	Included participants and clinical setting match the review question.
High concern for applicability	Included participants and clinical setting were different from the review question.
Unclear concern for applicability	If relevant information about the participants is not reported.

Table 8. Predictors domain: guidance notes for rating risk of bias and applicability

Domain 2: Predictors	
Risk of bias assessment	
Background: Bias in model performance can occur when the definition and measurement of predictors is flawed. Predictors are the variables evaluated for their association with the outcome of interest. Bias can occur, for example when predictors are not defined in a similar way for all participants or knowledge of the outcome influences predictor assessments.	
2.1 Were predictors defined and assessed in a similar way for all participants?	
Yes/ Probably yes	If definitions of predictors and their assessment were similar for all participants.
No/ Probably no	If different definitions were used for the same predictor or if predictors requiring subjective interpretation were assessed by differently experienced assessors.
No information	If there is no information on how predictors were defined or assessed.
2.2 Were predictor assessments made without knowledge of outcome data?	
Yes/ Probably yes was	If outcome information was stated as not used during predictor assessment or clearly not available to those assessing predictors.
No/ Probably no	If it is clear that outcome information was used when assessing predictors.
No information	No information on whether predictors were assessed without knowledge of outcome information.
2.3 Are all predictors available at the time the model is intended to be used?	
Yes/ Probably yes for	All included predictors would be available at the time the model would be used prediction.
No/ Probably no	Predictors would not be available at the time the model would be used for prediction.
No information	No information on whether predictors would be available at the time the model is intended to be used.
Risk of bias introduced by predictors or their assessment:	
Low risk of bias	If the answer to all signalling questions is “Yes” or “Probably Yes” then risk of bias can be considered low. If one or more of the answers is “No” or “Probably no”, the judgement could still be “Low risk of bias” but specific reasons should be provided why the risk of bias can be considered low, e.g. use of objective predictors not requiring subjective interpretation.
High risk of bias	If the answer to any of the signalling questions is “No” or “Probably no” there is a potential for bias.
Unclear risk of bias	If relevant information is missing for some of the signalling questions and none of the signalling questions is judged to put the domain at high risk of bias.

Domain 2: Predictors	
Concerns for applicability	
Background: The definition, assessment and timing of predictors in the primary study should be relevant to the review question, for example predictors should be measured using methods potentially applicable to the daily practice that is addressed by the review. **Concern that the definition, assessment or timing of predictors in the model do not match the review question:**	
Low concern for applicability	Definition, assessment and timing of predictors match the review question.
High concern for applicability	Definition, assessment or timing of predictors was different from the review question.
Unclear concern for applicability	If relevant information about the predictors is not reported.

Table 9. Outcome domain: guidance notes for rating risk of bias and applicability

Domain 3: Outcome	
Risk of bias assessment	
Background: Bias in model performance can occur when methods used to determine outcomes incorrectly classify participants with or without the outcome. Bias in methods of outcome determination can result from use of suboptimal methods, tests or criteria that lead to unacceptably high levels of errors in outcome determination, when methods are inconsistently applied across participants, and when knowledge of predictors influence outcome determination. Incorrect timing of outcome determination can also result in bias.	
3.1 Was the outcome determined appropriately?	
Yes/ Probably yes	If a method of outcome determination has been used which is considered optimal or acceptable by guidelines or previous publications on the topic. Note: This is about level of measurement error within the method of determining outcome (see concerns for applicability about whether the definition of the outcome method is appropriate).
No/ Probably no	If a clearly suboptimal method has been used that causes unacceptable error in determining outcome status in participants.
No information	No information on how outcome was determined.
3.2 Was a pre-specified or standard outcome definition used?	
Yes/ Probably yes	If the method of outcome determination is objective *or* if a standard outcome definition is used *or* if pre-specified categories are used to group outcomes.
No/ Probably no	If the outcome definition was not standard and not pre-specified.
No information	No information on whether the outcome definition was pre-specified or standard.
3.3 Were predictors excluded from the outcome definition?	
Yes/ Probably yes	If none of the predictors are included in the outcome definition.
No/ Probably no	If one or more of the predictors forms part of the outcome definition.
No information	No information on whether predictors are excluded from the outcome definition.
3.4 Was the outcome defined and determined in a similar way for all participants?	
Yes/ Probably yes	If outcomes were defined and determined in a similar way for all participants.
No/ Probably no	If outcomes were clearly defined and determined in a different way for some participants.
No information way.	No information on whether outcomes were defined or determined in a similar for all participants.
3.5 Was the outcome determined without knowledge of predictor information?	
Yes/ Probably yes knowledge of	If predictor information was not known when determining the outcome status, *or* outcome status determination is clearly reported as determined without predictor information.
No/ Probably no	If it is clear that predictor information was used when determining the outcome status.
No information	No information on whether outcome was determined without knowledge of predictor information.
3.6 Was the time interval between predictor assessment and outcome determination appropriate?	
Yes/ Probably yes relevant	If the time interval between predictor assessment and outcome determination was appropriate to enable the correct type and representative number of outcomes to be recorded, *or* if no information on the time interval is required to allow a representative number of the relevant outcome occur or if predictor assessment and outcome determination were from samples or information taken within an appropriate time interval.

Domain 3: Outcome	
No/ Probably no relevant	If the time interval between predictor assessment and outcome determination is too short or too long to enable the correct type and representative number of outcomes to be recorded.
No information	If no information was provided on the time interval between predictor assessment and outcome determination.
Risk of bias introduced by predictors or their assessment:	
Low risk of bias	If the answer to all signalling questions is "Yes" or "Probably yes" then risk of bias can be considered low. If one or more of the answers is "No" or "Probably no", the judgement could still be low risk of bias, but specific reasons should be provided why the risk of bias can be considered low, e.g. when the outcome was determined with knowledge of predictor information but the outcome assessment did not require much interpretation by the assessor (e.g. death regardless of cause).
High risk of bias	If the answer to any of the signalling questions is "No" or "Probably no" there is a potential for bias.
Unclear risk of bias	If relevant information about the outcome is missing for some of the signalling questions and none of the signalling questions is judged to put this domain at high risk of bias.
Concerns for applicability	
Background: The definition of outcome in the primary study should be relevant for the outcome definition in the review question.	
Concern that the outcome definition, timing or determination do not match the review question:	
Low concern for applicability	Outcome definition, timing and method of determination defines the outcome as intended by the review question.
High concern for another applicability	Choice of outcome definition, timing and method or determination defines outcome as intended by the review question.
Unclear concern for applicability	If relevant information about the outcome, timing and method of determination is not reported.

Table 10. Analysis domain: guidance notes for rating risk of bias

Domain 4: Analysis	
Risk of bias assessment	
Background: Statistical analysis is a critical part of prediction model development and validation. The use of inappropriate statistical analysis methods increases the potential for bias in reported model performance measures. Model development studies include many steps where flawed methods can distort results. We recommend reviewers seek statistical advice when completing assessments of the analysis domain.	
4.1 Were there a reasonable number of participants with the outcome?	
Yes/ Probably yes	For model development studies, if the number of participants with the outcome relative to the number of candidate predictor parameters is 20 or more (EPV ≥ 20).*
	For model validation studies, if the number of participants with the outcome is 100 or more.
No/ Probably no	For model development studies, the number of participants with the outcome relative to the number of candidate predictor parameters is less than 10 (EPV< 10).*
	For model validation studies, if the number of participants with the outcome is less than 100.
No information	For model development studies, no information on the number of candidate predictor parameters or number of participants with the outcome, such that the EPV cannot be calculated.
	For model validation studies, no information on the number of participants with the outcome.
* For EPVs between 10 and 20 the item should be rated as either probably yes or probably no, depending on the outcome frequency, overall model performance, and distribution of the predictors in the model. For more guidance see these references: (144-146)	
4.2 Were continuous and categorical predictors handled appropriately?	
Yes/ Probably yes	If continuous predictors are not converted into two or more categories when included in the model (i.e. dichotomised or categorised),
	or if continuous predictors are examined for nonlinearity using, for example, fractional polynomials or restricted cubic splines
	or if categorical predictor groups are defined using a pre-specified method.
No/ Probably no	If categorical predictor groups definitions do not use a pre-specified method.
	For model development studies, if continuous predictors are converted into two or more categories when included in the model.
	For model validation studies, if continuous predictors or categorical variables are categorised using different cut-points compared to the development study.
No information	No information on whether continuous predictors are examined for non-linearity.
	No information on how categorical predictor groups are defined,
	or no information on whether the same cut-points are used in the validation as compared to the development study.
4.3 Were all enrolled participants included in the analysis?	
Yes/ Probably yes	If all participants enrolled in the study are included in the data analysis.
No/ Probably no	If some or a subgroup of participants are inappropriately excluded from the analysis
No information	No information on whether all enrolled participants are included in the analysis.
4.4 Were participants with missing data handled appropriately?	

Domain 4: Analysis	
Yes/ Probably yes	If there are no missing values of predictors or outcomes *and* the study explicitly reports that participants are not excluded on the basis of missing data, *or* if missing values are handled using multiple imputation.
No/ Probably no	If participants with missing data are omitted from the analysis, *or* if the method of handling missing data is clearly flawed e.g. missing indicator method or inappropriate use of last value carried forward, *or* if the study had no explicit mention of methods to handle missing data.
No information	If there is insufficient information to determine if the method of handling missing data is appropriate.
4.5 Was selection of predictors based on univariable analysis avoided?	*[Development only]*
Yes/ Probably yes	If the predictors are *not* selected based on univariable analysis prior to multivariable modelling.
No/ Probably no	If the predictors are selected based on univariable analysis prior to multivariable modelling.
No information	If there is insufficient information to indicate that univariable selection is avoided.
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?	
Yes/ Probably yes	If any complexities in the data are accounted for appropriately, *or* if it is clear that any potential data complexities have been identified appropriately as unimportant.
No/ Probably no	If complexities in the data that could affect model performance are ignored.
No information	No information is provided on whether complexities in the data are present or accounted for appropriately if present.
4.7 Were relevant model performance measures evaluated appropriately?	
Yes/ Probably yes	If both calibration and discrimination are evaluated appropriately (including relevant measures tailored for models predicting survival outcomes)
No/ Probably no	If both calibration and discrimination are not evaluated, *or* if only goodness-of-fit tests, such as the Hosmer-Lemeshow test are used to evaluate calibration, *or* if for models predicting survival outcomes performance measures accounting for censoring are not used, *or* if classification measures (like sensitivity, specificity or predictive values) were presented using predicted probability thresholds derived from the dataset at hand.
No information	Either calibration or discrimination are not reported, *or* no information is provided as to whether appropriate performance measures for survival outcomes are used (e.g. references to relevant literature or specific mention of methods such as using Kaplan-Meier estimates) *or* no information on thresholds for estimating classification measures is given.
4.8 Was model overfitting and optimism in model performance accounted for?	*[Development only]*
Yes/ Probably yes	If internal validation techniques, such as bootstrapping and cross-validation have been used to account for any optimism in model fitting, and subsequent adjustment of the prediction model performance and presented model parameters have been applied.
No/ Probably no	If no internal validation has been performed, *or* if internal validation consists only of a single random split-sample of participant data,

Domain 4: Analysis	
	or if the bootstrapping or cross-validation did not include all model development procedures including any variable selection
No information	No information is provided on whether all model development procedures are included in the internal validation techniques.
4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? *[Development only]*	
Yes/ Probably yes	If the predictors and regression coefficients in the final model correspond to reported results from multivariable analysis.
No/ Probably no	If the predictors and regression coefficients in the final model do not correspond to reported results from multivariable analysis.
No information	If it is unclear whether the regression coefficients in the final model correspond to reported results from multivariable analysis.
Risk of bias introduced by the analysis:	
Low risk of bias	If the answer to all signalling questions is "Yes" or "Probably yes" then risk of bias can be considered low. If one or more of the answers is "No" or "Probably no", the judgement could still be low risk of bias, but specific reasons should be provided why the risk of bias can be considered low.
High risk of bias	If the answer to any of the signalling questions is "No" or "Probably no" there is a potential for bias.
Unclear risk of bias	If relevant information about the analysis is missing for some of the signalling questions but none of the signalling question answers is judged to put the analysis at high risk of bias.

Table 11. Overall assessment of risk of bias and concerns for applicability

Reaching an overall judgement of risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model evaluation can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement of concerns for applicability of the prediction model evaluation	
Low concerns for applicability	If low concerns for applicability for all domains, the prediction model evaluation is judged to have **low concerns for applicability**.
High concerns for applicability	If high concerns for applicability for at least one domain, the prediction model evaluation is judged to have **high concerns for applicability**.
Unclear concerns for applicability	If unclear concerns (but no "high concern") for applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns for applicability** overall.

Table 12. Suggested Tabular Presentation for PROBAST Results

Study	Risk of bias				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Risk of bias	Applicability
Study 1	+	-	?	+	+	+	+	-	+
Study 2	+	+	+	+	+	+	+	+	+
Study 3	+	+	+	?	-	+	+	?	-
Study 4	-	?	?	-	+	+	-	-	-
Study 5	+	+	+	+	+	?	+	+	?
Study 6	+	+	+	+	?	+	?	+	?
Study 7	?	?	+	?	+	+	+	?	+
Study 8	+	+	+	+	+	+	+	+	+

Box 1. Types of diagnostic and prognostic modelling studies or reports addressed by PROBAST
(adopted from the TRIPOD and CHARMS guidance(8, 19))

Prediction model development without external validation
These studies aim to develop one or more prognostic or diagnostic prediction models from a specific development data set. They aim to identify the important predictors of the outcome under study, assign weights (e.g. regression coefficients) to each predictor using some form of multivariable analysis, develop a prediction model to be used for individualised predictions, and quantify the predictive performance of that model in the development set. Sometimes, model development studies may also focus on adding one or more new predictors to established predictors. In any prediction model study, overfitting may occur, particularly in small data sets. Hence, development studies should include some form of resampling or "internal validation" (internal because the same data are used for both development and internal validation), such as bootstrapping or cross-validation. These methods quantify any optimism (bias) in the predictive performance of the developed model.

Prediction model development with external validation
Studies that have the same aim as the previous type, but the development of the model is followed by quantifying the model predictive performance in data *external* to the development sample i.e. from different participants. This may be data collected by the same investigators, commonly using the same predictor and outcome definitions and measurements, but sampled from a later time period (temporal validation); by other investigators in another hospital or country, sometimes using different definitions and measurements (geographic validation); in similar participants, but from an intentionally chosen different setting (e.g. model developed in secondary care and tested in similar participants from primary care); or even in other types of participants (e.g. model developed in adults and tested in children). Randomly splitting a single data set into a development and a validation data set is often erroneously referred to as a form of external validation, but actually is an inefficient form of "internal" validation, because the two so created data sets only differ by chance and sample size of model development is reduced**.**
When a model predicts poorly when validated in other data, a model validation can be followed by adjusting (or updating the existing model (e.g. by recalibration of the baseline risk or hazard or adjusting the weights of the predictors in the model) to the validation data set at hand, and even by extending the model by adding new predictors to the existing model. In both situations in fact a new model is being developed after the external validation of the existing model.

Prediction model external validation
These studies aim to assess the predictive performance of one or more existing prediction models by using in data *external* to the development sample i.e. from different participants.

Box 2. Differences between diagnostic and prognostic prediction model studies

Diagnostic prediction models aim to estimate the probability that a target condition measured using a reference standard (referred to as outcome in PROBAST) is currently present or absent within an individual. In diagnostic prediction model studies, the prediction is for an outcome already present so the preferred design is a cross-sectional study although sometimes follow-up is used as part of the reference test to determine the target condition presence at the moment of prediction.

Prognostic prediction models estimate whether an individual will experience a specific event or outcome in the future within a certain time period, ranging from minutes to hours, days, weeks, months or years: always a longitudinal relationship.

Despite the different timing of the predicted outcome, there are many similarities between diagnostic and prognostic prediction models, including the:

- Type of outcome is often binary (target condition or disease presence (yes/no) or future occurrence of an outcome event (yes/no).
- Key interest is to estimate the probability of an outcome being present or occurring in the future based on multiple predictors with the purpose of informing individuals and guiding decision-making.
- Same challenges occur when developing or validating multivariable prediction models. The same measures for assessing predictive performance of the model can be used, although diagnostic models more frequently extend assessment of predictive performance to focus on thresholds of clinical relevance.

There are also various differences in terminology between diagnostic and prognostic model studies:

Diagnostic prediction model study	**Prognostic prediction model study**
Predictors	
Diagnostic tests or index tests	Prognostic factors or prognostic indicators
Outcome	
Reference standard used to assess or verify presence/absence of target condition	Event (future occurrence yes or no) Event measurement
Missing outcome assessment	
Partial verification, lost to follow-up	Lost to follow-up and censoring

Box 3. Examples of systematic review questions for which PROBAST is suitable

There are various different questions that systematic reviews of prediction models may address. The following are examples of different types of review in which PROBAST can be applied.

A specific target population

- Review of all models developed or validated for predicting the risk of incident type 2 diabetes in the general population.(58)
- Review of all prognostic models developed or validated for use in patients diagnosed with acute stroke.(59)

A specific outcome

- Review of all diagnostic models developed or validated for detecting venous thromboembolism regardless the type of patients.(60)
- Review of all prognostic models developed or validated for predicting loss of daily activity, regardless the type of patients.(61)

A particular clinical field:

- Review of all prognostic models developed or validated in reproductive medicine.(62)
- Review of all prognostic models developed or validated in acute care of traumatic brain injury.(63)

A specific prediction model:

- Review of the predictive performance of the EuroSCORE (a model to predict operative mortality following cardiac surgery) as found across all external validation studies of the EuroSCORE model.(64)
- Review to compare the predictive performance of various prognostic models for developing cardiovascular disease in middle aged individuals in the general populations, across all validation studies of these models.(65)

A specific predictor:

- Meta-analysis of the added predictive value of C-reactive protein when added to the Framingham risk model.(66)
- Meta-analysis of the added predictive value of carotid artery imaging to an existing cardiovascular risk prediction model.(67)

Box 4. Prediction model performance measures

Calibration reflects the agreement between predictions from the model and observed outcomes. Calibration is preferably reported graphically, with observed risks plotted on the *y*-axis against predicted risks on the *x*-axis. This plot is commonly done by tenths of the predicted risk and is preferably augmented by a smoothed (lowess) line over the entire predicted probability range. This is possible both for prediction models developed by logistic regression(49, 76, 77) and by survival modelling(78, 79). The calibration plot displays the direction and magnitude of any model miscalibration across the entire predicted probability range, which can be combined with estimates of the calibration slope and intercept.(79, 80) Calibration is frequently assessed by calculating the Hosmer-Lemeshow goodness-of-fit test, however, this test has limited suitability to evaluate poor calibration and is sensitive to the numbers of groups and sample size: the test is often non-significant for small datasets and nearly always significant for large datasets. Studies reporting only the Hosmer-Lemeshow test with no calibration plot or a table comparing the predicted versus observed outcome frequencies provide no useful information on the accuracy of the predicted risks (see signalling question 4.7).

Discrimination refers to the ability of a prediction model to distinguish between individuals who do or do not experience the outcome event. The most general and widely reported measure of discrimination, for both logistic and survival models, is the concordance index (c-index), which is equivalent to the area under the receiver operating characteristic curve for logistic regression models.

Calibration and **discrimination** measures should take into account the type of outcome being predicted. For survival models, researchers should appropriately account for time-to-event and censoring, e.g. Harrell's c-index, D statistic.(81-83)

Many other model predictive performance measures are available including measures to express model classification abilities such as sensitivity, specificity and reclassification (e.g. the Net Reclassification Index) parameters.(80) These measures can be estimated after introducing one (or more) thresholds in the range of the model estimated probabilities. Classification measures are frequently used in diagnostic test accuracy studies but less in prediction model studies. Categorization of the predicted probabilities in two or more probability categories for estimation of classification measures can lead to loss of information, since the entire range of predicted probabilities of the model is not fully utilised. Using thresholds can allow discrimination to be reported at potentially clinically relevant thresholds as opposed to across all potential thresholds which may not be clinically important. However, introducing probability thresholds implies that the chosen threshold is relevant to clinical practice which often is not the case since these thresholds are often data driven yielding biased classification parameters.(84) Authors should rather assess these measures based on the general principles of pre-specifying (probability) thresholds (see also signalling question 4.2) to avoid multiple testing of thresholds and potential selective reporting of thresholds based on the data itself.

There are many other measures of performance measure including net benefit measures and decision curve analysis.(85) Many of these measures provide a link between probability thresholds and false-positive and false-negative results to obtain the model net benefit at a particular threshold. Net benefit measures are not commonly reported for prediction modelling studies.

All the above model performance measures, when estimated on the development data, are often optimistic due to overfitting or choosing optimal thresholds, and should therefore be estimated using bootstrapping or cross-validation methods (see signalling question 4.8).

PROBAST: a tool to assess risk of bias and applicability of prediction model studies – explanation and elaboration

Figure

Figure 1. Suggested Graphical Presentation for PROBAST Results